RoD-TAL: A Benchmark for Answering Questions in Romanian Driving License Exams

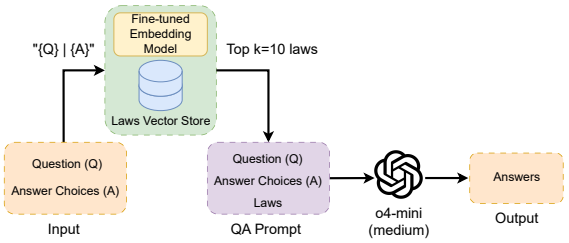
Anonymous ACL submission

Abstract

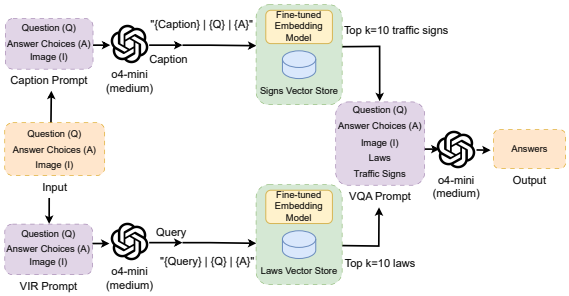
The intersection of AI and legal systems presents a growing need for tools that support legal education, particularly in under-resourced languages such as Romanian. In this work, we aim to evaluate the capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs) in understanding and reasoning about Romanian driving law through textual and visual question-answering tasks. To facilitate this, we introduce RoD-TAL, a novel multimodal dataset comprising Romanian driving test questions, text-based and image-based, alongside annotated legal references and explanations written by human experts. We implement and assess retrieval-augmented generation (RAG) pipelines, dense retrievers, and reasoning-optimized models across tasks, including Information Retrieval (IR), Question Answering (QA), Visual IR, and Visual QA. Our experiments demonstrate that domain-specific fine-tuning significantly enhances retrieval performance. At the same time, chain-of-thought prompting and specialized reasoning models improve QA accuracy, surpassing the minimum grades required to pass driving exams. We highlight the potential and the limitations of applying LLMs and VLMs to legal education.

1 Introduction

The intersection of Artificial Intelligence (AI), legal systems, and Web technologies offers a powerful avenue for enhancing public access to structured legal knowledge. Road traffic law, in particular, provides a rule-based, codified domain that is well-suited for computational reasoning and the development of intelligent, web-based educational tools. As web information systems evolve to integrate data-driven AI models, legal education, especially in underserved linguistic and regional contexts, remains a vital yet underexplored application area (Lai et al., 2024).



(a) QA pipeline



(b) VQA pipeline

Figure 1: Architectures of the QA and VQA pipelines.

Despite significant advances in Large Language Models (LLMs) and retrieval-augmented architectures, their application to legally grounded question answering in low-resource languages remains limited (Hijazi et al., 2024; Das et al., 2024). In countries like Romania, where access to legal interpretation and educational resources is often inconsistent, there is a growing need for inclusive, intelligent systems that can support legal literacy and public understanding of the law (Guha et al., 2023; Hoppe et al., 2021).

Consequently, we propose the following objectives for this research. (1) Evaluate LLMs and their reasoning capability on question answering (QA) and visual QA (VQA) tasks in the legal and driving settings. (2) Assess models' performance, biases, and limitations. (3) Investigate how reliably these models can be integrated to support education and law-related tasks.

We curated and created a novel dataset, called

Work/Author	Backbone	Modality	Accuracy				
			US	JP	CN	SG	RO
Zero-Shot QA (Zhou et al., 2024b)	GPT-4	Text	92.1%	86.5%	85.2%	88%	-
Zero-Shot VQA (Zhou et al., 2024a)	GPT-4V	Image	-	66.7%	79.2%	-	-
IDKB (Lu et al., 2025)	GPT-4o	Image	Overall for 15 languages: 53%				
RAG QA (Ours - Fig. 1a)	o4-mini	Text	-	-	-	-	86.4%
RAG VQA (Ours - Fig. 1b)	o4-mini	Image	-	-	-	-	78.3%

Table 1: Comparison of models and accuracy over languages. Results are reported for both text and image modalities, covering multiple languages (US - United States of America; JP - Japanese; CN - Chinese; SG - Singapore, RO - Romanian), with our method evaluated on Romanian for both modalities.

RoD-TAL, composed of **Ro**manian **D**riving **T**ests **a**nd **L**aws (annotated and referenced in the QA/VQA pairs), to help us assess our objectives.

The main contributions of this work are:

- Introduce a novel dataset called RoD-TAL.
- Evaluate LLMs in the context of a low-resource language, Romanian, and a legal domain setting, exposing their biases and limitations.
- Propose a foundation for a legal Romanian-based dense retriever that doesn't focus solely on semantic similarity but on question and legal documents alignments.
- Propose solutions (see Figure 1 and Table 1) for all our identified research areas: information retrieval (IR), question answering (QA), visual IR (VIR), and visual QA (VQA), with strong results across all tasks.
- Present a thorough analysis of hallucinations in legal QA settings (Appendix D), ranging from citations, instruction following, and failure patterns.

2 Related Work

Driving QA and VQA. Zhou et al. (2024b) evaluated ChatGPT (GPT-4) (OpenAI et al., 2024b) on 814 written driving-license questions from California, Tokyo, Beijing, and Singapore, assessing performance across dimensions such as legal reasoning, situational understanding, and safety bias. Accuracy ranged from 85.2% to 92.1%, with lower scores on region-specific legal questions (e.g., 63.2% in China). While the model performed well overall, it showed limitations in handling local regulations and context-sensitive reasoning.

In follow-up work, Zhou et al. (2024a) evaluated Vision-Language Models (VLMs) such as ChatGPT and Bard on visual driving-license questions from Tokyo and Beijing. While models performed moderately on traffic sign recognition (70%) and better on scenario-based (80%) and combined visual tasks (80%), the study highlighted ongoing challenges in applying VLMs to real-world autonomous driving contexts.

However, these studies evaluated out-of-the-box LLM or VLM performance; they did not aim to optimize model accuracy and did not incorporate retrieval augmentation techniques. While GPT models have likely encountered many laws and question pairs during pretraining, reframing this task as a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) problem could offer a more principled and scalable approach to improving legal and regulation-specific reasoning.

Vision-Language Driving Datasets. Multiple works (Kim et al., 2018; Deruyttere et al., 2019; Li et al., 2022; Malla et al., 2023; Qian et al., 2024; Sima et al., 2024; Shao et al., 2024; Tian et al., 2024; Park et al., 2024; Lu et al., 2025) have introduced and evaluated vision language driving datasets for autonomous driving, some based on open QA, or Multiple-Choice Question Answering (MCQA), and spanning multiple languages, topics, and vehicle categories; see Table 2 for a comparison of these resources. Additionally, their limitation lies in not having references to legal corpora and relying solely on the internal knowledge of vision language models during QA tasks.

Multiple-Choice QA with LLMs Across Domains. Zhong et al. (2024) assessed LLMs, including GPT-4, on Chinese-based human exams like the Law School Admission Test (LSAT) and Lawyer Qualification Test (LQT), using formats such as multiple-choice and fill-in-the-blank. Among various prompting strategies, Few-Shot with chain-of-thought (CoT) performed best. GPT-4 scored 34–40% on the LQT and 31–87% on the LSAT. The study found stronger reasoning in high-resource languages (e.g., English) and domains like history and logic, while performance was weaker in law, math, and physics. Challenges included concept disambiguation, strict logical reasoning, and multi-

Dataset	Data Type		Data Source		Data Domain			Knowledge Domain				Size	Law Ref.
	QA	MCQ	Real	Synth.	Country	Lang.	Type	LR	SS	DT	DD		
BDD-X (Kim et al., 2018)	✓	✗	✓	✗	US	EN	Car	✗	✓	✗	✗	26K	✗
Talk2Car (Deruyttere et al., 2019)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	12K	✗
CODA-LM (Li et al., 2022)	✗	✓	✗	✓	DE, CN, SG	EN	Car	✗	✓	✗	✗	10K	✗
DRAMA (Malla et al., 2023)	✗	✗	✓	✗	JP	EN	Car	✗	✓	✗	✗	102K	✗
nuScenes-QA (Qian et al., 2024)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	460K	✗
DriveLM (Sima et al., 2024)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	2M	✗
LangAuto CARLA (Shao et al., 2024)	✓	✗	✗	✓	US	EN	Car	✗	✓	✗	✗	64K	✗
SUP-AD (Tian et al., 2024)	✗	✓	✓	✗	CN	EN	Car	✗	✓	✗	✗	-	✗
VLAAD (Park et al., 2024)	✗	✓	✓	✗	US	EN	Car	✗	✓	✗	✗	64K	✗
IDKB (Lu et al., 2025)	✓	✓	✓	✓	15	9	4	✓	✓	✓	✓	1M	✗
RoD-TAL (Ours)	✗	✓	✓	✗	RO	RO	Car	✓	✓	✓	✓	1.2K*	✓

Table 2: Comparison of datasets, domains, and knowledge coverage, inspired by Lu et al. (2025). The data source can be real or synthetic (synth.). Knowledge Domain spans Laws and Regulation (LL), Signs and Signals (SS), Driving Techniques (DT), and Defense Driving (DD). *RoD-TAL spans 1.2K questions, out of which 400 contain images.

hop inference, highlighting areas for further improvement. In the context of Romanian works, there has been a proposed MCQA dataset for the legal domain in Craciun et al. (2025) curated from legal examinations from different levels and specializations, as well as an MCQA dataset for the medical domain (Dima et al., 2024), built from university entrance examinations.

Multiple-Choice VQA. Das et al. (2024) proposed EXAMS-V, a multilingual, multimodal benchmark for evaluating VLMs on multiple-choice questions across diverse domains (excluding law). GPT-4V scored an average of 42.5%, revealing limitations in multimodal reasoning and the integration of visual and textual information, despite showing early potential.

Information Retrieval and Retriever Fine-Tuning. Moreira et al. (2024) investigated fine-tuning dense retrievers for RAG tasks, emphasizing hard-negative mining strategies. Their results showed that starting from positive-aware setups and gradually introducing harder negatives significantly improved retrieval accuracy and response quality, underscoring the importance of training data curation in RAG pipelines.

Visual Information Retrieval. Dong et al. (2024) proposed a modality-aware retrieval approach that leverages visual LLMs to generate dense image captions for use in downstream querying, enhancing retrieval performance by integrating visual content more effectively.

3 RoD-TAL: Romanian Driving Tests And Laws

The RoD-TAL novel resource comprises a law corpus from the Romanian legislation, which we refer to as RoD-Law, and the QA part, including text and image-based questions, referred to as RoD-QA.

Corpus Source	Articles
Traffic Regulation Rules	225
Road Code	147
Penal Code (traffic-related)	9
Technical Inspection Law	15
Civil Auto Liability Insurance Law	47
Traffic Signs	140

Table 3: Distribution of articles in the RoD-Law corpus by legal source.

Dataset	Modality	Law Ref.	Size
Split 1	Text-based	✓	638
Split 2	Text-based	✗	131
Split 3	Image-based	✓	316
Split 4	Image-based	✗	71

Table 4: RoD-QA splits by modality and law reference annotation.

3.1 RoD-Law Corpus

A central component of the RoD-TAL framework is its law-grounded foundation, the RoD-Law corpus, which ensures every answer can be explicitly traced back to the Romanian legislation. The legal corpus, curated as the retrieval base for all downstream tasks, was compiled from official sources, with abrogated sections removed to ensure relevance and currency, until March 2025. The composition of the legal corpus is summarized in Table 3, including laws from the traffic regulations, road code, penal code, and technical inspection law, and civil auto liability insurance law.

The RoD-Law legal corpus is complemented by an annotated collection of 140 distinct traffic signs, extracted from the answer references across the dataset and spanning 11 major categories (see

Figure 4. Each sign is provided with its name, category, and a concise explanation, supporting both QA annotation and VQA tasks. Statistics are also presented in Appendix A

3.2 RoD-QA Dataset

Dataset Structure. Built upon the curated legal base, the RoD-TAL dataset consists of multiple-choice questions sourced from the Romanian driving-license written tests, henceforth referred to as RoD-Law, available on the public educational platform **Scoala Rutiera**[1]. Each question is annotated with relevant legal references from RoD-Law, allowing for the evaluation of both standard LLM answering and retrieval-augmented generation (RAG) systems grounded in actual law. *Scoala Rutiera* is one of the largest platforms in Romania for learning and preparing for theoretical driving tests, according to internet traffic[2] and mobile app downloads on Google Play[3]. As claimed on their website, the questions are the official ones provided by the authorities during the exams. Therefore, we use this assumption to consider that these annotations are correct.

The data structure for each sample includes the question, a set of candidate answers (with an explicit correct answer or more), explanation, legal reference, and a list of indicators where relevant. Visual questions were further categorized and enriched by o4-mini-based (OpenAI, 2025) sign annotation, followed by manual verification. For the experimental setup, we evaluate various combinations of text and image-based questions with or without law annotations. An overview of the dataset splits is provided in Table 4.

This comprehensive annotation schema supports evaluation not only of LLM answer accuracy but also of legal retrieval precision/recall and legal grounding in RAG setups. The presence of visual (VQA) questions further enables research into legally grounded multimodal models, an emerging area in AI and law.

Dataset Statistics. The dataset encompasses 18 question categories in total, with the visual subset categorized into three additional secondary categories: point of view (pov), aerial, and miscellaneous (misc). Figure 3 details the distribution among visual-question secondary categories. Figure 7 from the Appendix A shows the number of questions per primary category. We further provide a fine-grained breakdown of questions by primary category and dataset split in Figure 2. Some categories are better represented than others depending on the split. For example, split 1 contains more questions related to *driver obligations* and *sanctions and offenses*, while the same categories contain fewer samples in the rest of the splits. For more statistics, refer to Appendix A.

4 Experiments

The problem of developing an AI system capable of answering legal questions based on Romanian traffic law can be decomposed into several distinct but interconnected tasks. These tasks span both textual and visual modalities, collectively defining the core components required for building, evaluating, and improving such a system. By segmenting the challenge into modular tasks, we facilitate targeted experimentation, fine-grained performance evaluation, and the possibility of optimizing each sub-component independently. We can pursue four topics: Information Retrieval (split 1), Question Answering (splits 1 and 2), Visual Information Retrieval (split 3), and Visual Question Answering (splits 3 and 4). We make our code publicly available on the GitHub repository[4].

4.1 Information Retrieval

An essential component of our system is the ability to retrieve relevant legal text passages that justify the correct answer to a question. Our focus is on optimizing Recall@k while minimizing k, since high values are impractical for downstream processing and the LLM generation part. Based on the distribution of legal articles that ground the questions (Figure 8 from Appendix A), we consider $k = 10$ sufficient for our experiments such that the LLM is context-bloated downstream.

In our work, we experiment with the embedding model $\text{mE5}_{\text{small}}$ (Wang et al., 2024), where we evaluate different query building techniques, rerankers, or fine-tuning on real or augmented data. The full experimental details can be found in Appendix B.5.

4.2 Question Answering

We utilize a CoT prompting strategy with the GPT-4o mini model, incorporating 10 retrieved docu-

[1] https://www.scoalarutiera.ro/

[2] https://www.similarweb.com/website/scoalarutiera.ro/#overview

[3] https://play.google.com/store/apps/details?id=com.app.scoalarutiera

[4] https://anonymous.4open.science/r/rodtal-v2-F50C/

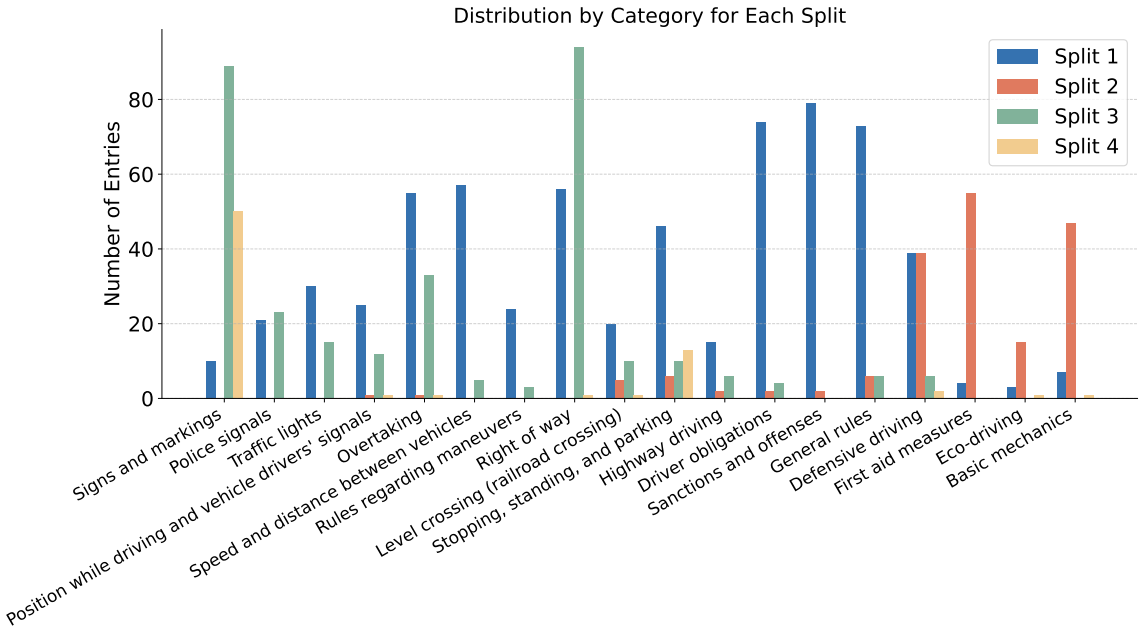


Figure 2: Number of questions per primary category and split.

ments alongside the question and answer options in the context. This setup reflects a standard RAG pipeline. To assess the value of retrieval, we experiment with a baseline prompt-based approach that does not employ document retrieval. This allows us to isolate the benefit of retrieval in performance. Additionally, we evaluate an *ideal RAG* setup, in which only the exact relevant documents are included, simulating the ideal recall and precision conditions. We explore prompt engineering based on observed failure patterns in the model's behavior and propose a better prompt (henceforth denoted BP). We also conduct experiments without the CoT technique to test the role of step-by-step reasoning in answer accuracy. Lastly, we experiment with reasoning-tuned models under the same input specifications and prompts. We explore reasoning models (o4-mini) (OpenAI, 2025) with the same specifications and prompts as before to see if native reasoning helps with the problem. We also assess the performance of open-weights models, such as Mistral Small 3.1 Instruct[5] and Gemma 3 27B Instruct (Team et al., 2025), following the same scenarios as with the GPT-4o mini (OpenAI et al., 2024a) model.

4.3 Visual Information Retrieval

In the VIR setting, we evaluate the performance of retrieving the correct laws based on the image and its associated multiple-choice question. We, separately, retrieve and identify the correct traffic signs in the presented image. We adopt several methodologies to enhance a text query with the image characteristics. As a baseline, we compare against a plain QA search. First, we generate and save 50-100-word captions using the o4-mini with CoT. We propose a combination of caption + QA. Then, we use the o4-mini model to rewrite our queries. We evaluate the following scenarios: Image + QA, Image + Caption + QA. For an extra measure, we also add the QA for the latter examples, in case the LLM omits some details For testing traffic sign retrieval, a similar methodology is followed. However, we test two embedding scenarios: first, we only embed the traffic sign name and category, and second, we also append the explanation.

4.4 Visual Question Answering

We assess the performance of CoT prompting on the o4-mini model in three main configurations:

- **Model's prior knowledge**: The model answers questions using just the question, candidate answers, and image, without any external retrieval.
- **Ideal RAG**: The model is provided with exactly the relevant documents or legal references for each question (i.e., perfect retrieval conditions).
- **Best RAG**: The model is provided with the best retrieval methodology from the VIR task for laws and traffic signs.

[5] https://mistral.ai/news/mistral-small-3-1

We also propose a similar methodology for the VIR task, comparing Caption + QA, Image + QA, and Image + Caption + QA, all leveraging CoT.

4.5 Evaluation Metrics

For information retrieval, we employ Recall@k, Precision@k, and nDCG@k (Järvelin and Kekäläinen, 2002). Recall@k measures the proportion of relevant documents among the top-k retrieved, while Precision@k evaluates the proportion of retrieved documents in the top-k that are relevant. nDCG@k further incorporates ranking position, assigning a higher weight to relevant documents appearing earlier in the list. Out of these, the most important are recall@k and nDCG@k, while precision@k can be misleading due to the retrieval of more documents than needed.

For QA and VQA tasks, we use Precision, Recall, F1 score, and Exact Match (EM). Precision and Recall assess prediction accuracy and completeness, while F1 balances both. EM is our primary metric, requiring an exact match with the ground truth and offering no partial credit. Given our emphasis on strict answer correctness, EM is the central measure of QA performance. Detailed experimental results for the IR/QA RAG setup can be found in Appendix E.

4.6 Experimental Setup

Details about the experimental setup are presented in Appendix B. All the metric comments refer to the exact match score, which is also the performance that candidates are evaluated on in the MCQA setting.

5 Results

5.1 Information Retrieval

Our baseline experiment (1) in Table 5 yields only a 50% Recall@10 performance. Answer choices increase this performance by almost 10% in (2), thus a lot of the information needed for search is also present in the answer choices. Retrieving 40 documents and re-ranking the top 10, we also see a boost in performance by 8%-10%. Using an LLM to rephrase the query (5) yields similar results as with the QA (2); thus, the problem is semantic and not syntactic.

By fine-tuning our embedding model on the dataset, and now looking only at the test split, we see an 18% improvement for (6) in the Recall metric. Adding back a Reranker (7) hinders our performance by 11% (observed on split 2 specifically). This has its roots in the misalignment from which the Retriver model also suffered. The augmented fine-tuned model only yields 3% performance over its baseline (see (8) compared to (2)) but underperforms massively compared to the one trained on the data. This suggests a distribution shift from the original one: the complexity of the questions is lower, and the amount of related articles (as stated earlier) is much smaller.

These findings emphasize the importance of domain-specific representation learning for legal NLP in under-resourced languages and demonstrate that fine-tuning on even modestly sized, targeted datasets can significantly improve performance, more so than generic improvements in model architecture or reranking strategies.

5.2 Question Answering

Our first three experiments (1), (2) and (3) in Table 6 aim to compare the Retrieval component. On the first split, we observe that the RAG component improves performance by 11%, and comparing the best setup with an ideal RAG, the difference is marginal. This shows that even with not all documents retrieved or some over-retrieved and not relevant, the LLM is capable of completing the missing information and responding in a similar manner (only 2% variation).

Error analysis revealed three main failure modes (see also Appendix F):

- **Difficult Questions**: Difficulty handling nuanced or misleading options.
- **Safety Bias**: Prioritizing safe answers over legally correct ones.
- **Overthinking**: Overextending reasoning beyond the immediate question scope.

With this refinement in mind, using a better prompt to mitigate these findings, we achieve another 9% improvement over the previous results of strategy (4). At this step, we also want to ablate the CoT, making the model respond directly. This shows a loss of 25% in performance.

Adding a reasoning model over the previous experiments, we gain another 19% performance increase in strategy (6). Ablating Retrieval also shows it is still relevant, where we lose 22% in strategy (7).

Experiments with Mistral show similar patterns but overall weaker results. One thing to note is that this model would sometimes not follow the instructions, either by answering with the responses

Method	Retrieval - Train			Retrieval - Test		
	R@10	P@10	nDCG@10	R@10	P@10	nDCG@10
(1) Question based (Q)	50.15	11.45	38.75	49.29	11.01	37.48
(2) Question + Answer Choices (QA)	60.05	14.00	51.33	59.31	13.43	51.23
(3) QA + ReRanker jina	68.70	16.47	64.15	70.08	16.71	63.90
(4) QA + ReRanker bert-msmarco	63.99	15.21	56.22	65.60	15.54	56.49
(5) QA rephrased using GPT 4o-mini	60.84	14.27	51.70	60.64	14.14	49.86
(6) Finetuned Retriever	**99.89**	**27.21**	**99.71**	**88.14**	**23.28**	**81.41**
(7) Finetuned Retriever + ReRanker jina	80.43	20.25	71.44	77.55	19.92	70.23
(8) Augmented Finetuned Retriever	63.80	14.80	57.10	62.76	14.53	57.43

Table 5: Precision@10, Recall@10, and nDCG@10 metric scores for the tried and tested methods on the Information Retrieval task.

Method	Split 1		Split 2
	Train	Test	Test
(1) GPT-4o mini + CoT + RAG	57.8	55.5	72.9
(2) GPT-4o mini + CoT w/o RAG	46.1	43.0	71.3
(3) GPT-4o mini + CoT + Ideal RAG	59.0	63.3	69.6
(4) GPT-4o mini + CoT + RAG + BP	67.8	75.0	79.6
(5) GPT-4o mini + RAG + BP w/o CoT	42.7	47.7	60.8
(6) o4-mini + CoT + RAG + BP	**86.3**	**91.4**	**83.4**
(7) o4-mini + CoT + BP w/o RAG	64.3	63.3	82.3
(8) Mistral + CoT + RAG	42.5	42.2	51.9
(9) Mistral + CoT w/o RAG	46.1	39.8	68.0
(10) Mistral + CoT + Ideal RAG	30.4	35.2	4.4
(11) Mistral + CoT + RAG + BP	47.6	53.9	48.1
(12) Mistral + RAG + BP w/o CoT	57.5	53.9	75.7
(13) Gemma 3 + CoT + RAG	60.8	53.9	75.1
(14) Gemma 3 + CoT w/o RAG	48.0	38.3	65.7
(15) Gemma 3 + CoT + Ideal RAG	61.0	57.0	71.8
(16) Gemma 3 + CoT + RAG + BP	67.1	53.1	80.7
(17) Gemma 3 + RAG + BP w/o CoT	55.1	46.9	72.4

Table 6: Exact match score on IR/QA RAG pipeline. BP indicates that a better prompt was used.

first and then arguing, or by starting an infinite loop of generating additional laws instead of answering. Also, for the experiments where we ablate CoT, the model would perform the same or better by 10-30%, depending on the dataset used.

Gemma achieves similar results compared to the OpenAI non-reasoning models. We discuss more about the comparison of open and closed source models, potential answer bias from our prompt and dataset difficulty in Appendix C.

5.3 Visual Information Retrieval

Using the fine-tuned retriever, our baseline text-based solution for retrieval yields a 60% recall, which is a strong initial result (strategy (1) in Table 7). Adding the caption to the search query improves the performance by 10%. Further adding the image information, by using an LLM to rephrase the query (entries denoted with `R[...]` represent rephrasing of the query using VLM), shows an improvement in the Image + QA scenario by 3% in recall and a loss of 3% when also using the caption. Contrary

Method	Retrieval Laws - Split 3		
	R@10	P@10	nDCG@10
(1) QA	60.45	12.72	56.18
(2) C + QA	70.30	15.06	56.41
(3) R[I + QA]	73.51	16.13	58.38
(4) R[I + C + QA]	67.70	14.87	52.85
(5) R[I + QA] + QA	**77.09**	**16.83**	**62.67**
(6) R[I + C + QA] + QA	75.17	16.32	59.67

Table 7: Precision@10, Recall@10, and nDCG@10 scores for the Visual Information Retrieval of Laws task. Notation: QA – question and answer choices, C – caption, I – image, `R[...]` – rephrasing using o4-mini.

to the belief that more context helps the model to produce a better query, it doesn't in this scenario. Concatenating the QA pair to the previous experiments also shows a 7-8% improvement on both scenarios, suggesting that some details were left out in the rephrasing part. Looking at retrieving traffic signs in Table 8, with the same experiments in mind, they follow a similar pattern in performance. When changing the embedding method to also include more context and details for the indicators (the entries which have a *), we can see a slight improvement in the Caption + QA scenario. This combination helps with matching extra details in the caption to the extra details in the description, seeing a 3% gain in strategy (2*) over the previous best.

5.4 Visual Question Answering

We aim to evaluate VQA based on three prompting scenarios and three retrieval scenarios (Table 9). First, we want to test the input combination as before, either Caption+QA, Image+QA, or Image+Caption+QA. Consistently, Image+QA has a better performance than Caption+QA by 7-8%, while Image+Caption+QA is only 2-3% above Caption+QA. This is a similar pattern we spotted in the retrieval task, where more context (the caption) doesn't improve the performance, but it does af-

Method	Retrieval Indicators - Split 3			Retrieval Indicators - Split 4		
	R@10	P@10	nDCG@10	R@10	P@10	nDCG@10
(1) QA	47.52	7.18	33.50	61.97	7.32	49.99
(2) C + QA	60.49	9.39	**46.62**	73.70	**9.29**	56.53
(3) R[I + QA]	60.39	9.20	42.93	70.18	8.59	49.84
(4) R[I + C + QA]	57.80	8.82	43.66	67.37	8.45	50.87
(5) R[I + QA] + QA	60.23	9.24	46.14	73.47	9.01	56.29
(6) R[I + C + QA] + QA	58.25	8.92	45.15	**74.41**	9.15	**60.38**
(1*) QA	41.98	6.13	28.72	64.08	7.60	48.38
(2*) C + QA	**63.13**	**9.65**	44.96	69.95	8.59	53.13
(3*) R[I + QA]	58.80	9.05	41.67	63.14	7.46	50.11
(4*) R[I + C + QA]	57.25	8.95	39.95	68.54	8.16	50.63
(5*) R[I + QA] + QA	58.93	8.98	41.93	70.89	8.59	54.89
(6*) R[I + C + QA] + QA	59.34	9.11	42.13	72.06	8.59	56.11

Table 8: Precision@10, Recall@10 and nDCG@10 metric scores the Visual Information Retrieval of Indicators task. Notation: QA – question and answer choices, I – image, C – caption, `R[...]` – rephrasing using o4-mini, * – corpus has different content embedded into vectors.

Method	Split 3	Split 4
(1) o4-mini + C + QA + CoT	64.2	71.8
(2) o4-mini + I + QA + CoT	71.5	74.6
(3) o4-mini + I + C + QA + CoT	64.9	74.6
(4) o4-mini + C + QA + CoT + Ideal RAG	69.9	78.9
(5) o4-mini + I + QA + CoT + Ideal RAG	**77.8**	78.9
(6) o4-mini + I + C + QA + Ideal RAG	71.8	78.9
(7) o4-mini + C + QA + CoT + RAG	67.4	76.1
(8) o4-mini + I + QA + CoT + RAG	75.6	**90.1**
(9) o4-mini + I + C + QA + RAG	69.3	77.5

Table 9: Exact match scores on the VIR/VQA RAG pipeline using strategy (5) for laws and strategy (2*) for indicators. Notation: QA – question and answer choices, I – image, C – caption.

fect performance. Checking on the Retrieval Component, we find the theoretical best and our best setup to be comparable in results (1-2% difference), while ablating RAG loses at least 5% performance.

5.5 Quantitative and Qualitative Analysis

We present some QA examples where the better prompting strategy improved the results and mitigated the initially observed limitations in Appendix F. We also present samples of wrong visual questions on each of the three secondary categories, using the best RAG setup.

We add a more in-depth error analysis in Appendix G, focusing on a detailed analysis of performance across all tasks per splits and categories. We examine the tendencies of over-selecting answers which is a direct factor to reducing Precision and Exact Match; and the number of reasoning steps used by LLMs which indicates us the categories of questions which are more difficult overall.

We discuss hallucinations in Appendix D. We analyze citation counts and show that the model tends to hallucinate documents when citing more documents in the same output, degrading performance at the same time. We analyze instruction following of the models and discover more failure rates for output formatting in Mistral 3.1, which leads to a performance downgrade due to not being able to parse outputs; and more failure rates for citing in the indicated way on OpenAI o4-mini, although maintaining superior performance. We analyze hallucinations in captions and show their direct correlations to performance downgrade. Finally, using LLM-as-a-Judge (Zheng et al., 2023) we find and then classify hallucinations, showing failure patterns in the legal context.

6 Conclusions

In this work, we evaluated Information Retrieval, Visual Information Retrieval, Question Answering, and Visual Question Answering both independently and in combination on a newly introduced dataset. Our results demonstrated promising performance across tasks, while also highlighting specific areas for improvement. In the IR component, further fine-tuning was needed to prevent the inclusion of actual positives during hard negative mining. For VIR, future work should explore more advanced methods incorporating joint image–text embeddings. In QA and VQA, improved prompting strategies or targeted model fine-tuning could help mitigate the limitations and biases inherent to current large language models.

Additionally, we examined the challenges LLMs face in multiple-choice QA settings, particularly the importance of minimizing over-selection to maximize precision. This research represents a novel contribution to the intersection of QA, VQA, IR, and VIR within the Romanian legal domain, specifically focused on traffic law, but with potential applicability to a wide range of legal tasks.

Limitations

We acknowledge that our work presents some limitations. The choice of not splitting larger documents into sub-chunks and truncating them instead aimed to simplify our experiments and benchmarking of the dataset, and we acknowledge that it might set us back from achieving a perfect retrieval result.

We also acknowledge that we didn't run the experiments with larger models, which could have improved the scores, due to hardware or budget constraints. For the same reasons, we didn't run the experiments on multiple seeds to explore variation.

Risks and Ethical Considerations

The only risk we see is in educational settings. It is known that LLMs suffer from hallucination (OpenAI et al., 2024a), which we also address in Appendix D. If users were to learn for the driving exams using our setup, we cannot guarantee the information is correct, and this is also reflected in the non-perfect results. In addition, our methods cannot be used for legal advice.

Since law may change in the future, and as discussed in Appendix B.5, the performance of the fine-tuned retriever model is not impacted if case law articles are repealed, by simply removing them from the corpus. We provide the code for scraping, cleaning, and normalizing articles from official sources, to benefit from the most recent changes in the law while using our experimental setup.

References

Cristian-George Craciun, Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, and Mihaela-Claudia Cercel. 2025. GRAF: Graph retrieval augmented by facts for Romanian legal multi-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12708–12742, Vienna, Austria. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. pages 7768–7791.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2Car: Taking control of your self-driving car. pages 2088–2098.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. pages 4531–4541.

Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. pages 2417–2429.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem Alzahrani, Hebah AlShamlan, George Turkiyyah, and Omar Knio. 2024. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. pages 225–249.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578.

Jacob Kohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024. Large language models in law: A survey. *AI Open*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, and 1 others. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European conference on computer vision*, pages 406–423. Springer.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. 2025. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5838–5846.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. pages 2014–2037.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. System Card, released April 16, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 980–987.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models, 2023. *URL https://arxiv. org/abs/2304.06364*.

Shanmin Zhou, Jialong Li, Takuto Yamauchi, Jinyu Cai, and Kenji Tei. 2024a. Evaluating vision-language models in visual comprehension for autonomous driving. In *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pages 205–209. IEEE.

Shanmin Zhou, Jialong Li, Mingyue Zhang, Daisuke Saito, Hironori Washizaki, and Kenji Tei. 2024b. Can chatgpt obey the traffic regulations? evaluating chatgpt's performance on driving-license written test. In

2024 International Conference on Innovation, Knowledge, and Management (ICIKM), pages 7–14. IEEE.

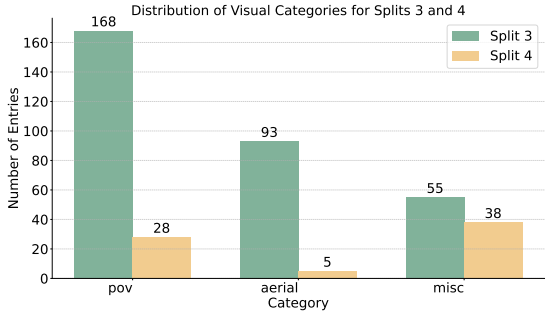


Figure 3: Number of visual questions per secondary category

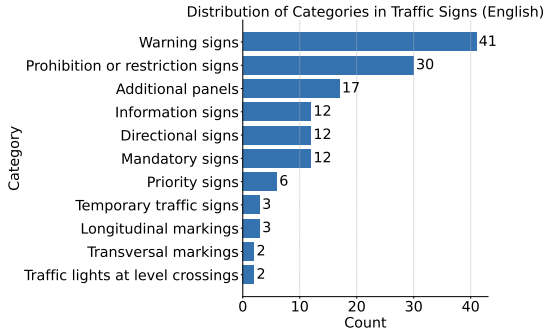


Figure 4: Distribution of traffic signs (indicators) per category.

A Dataset Statistics

In this section, we present statistics of the RoD-TAL dataset, which comprises two main components: the legal corpus RoD-Law and the multiple-choice question-answering dataset RoD-QA. We analyze the characteristics and distributions relevant to each component to provide insights into the dataset's structure and content.

A.1 RoD-Law Statistics

The corpus totals 443 legal documents from the Romanian law. In Figure 5, we present the distribution of the number of tokens from the entire RoD-Law corpus. The distribution roughly follows a long-tail power law, where most documents have fewer tokens, with approximately 84% of documents containing text under 500 tokens in length.

Out of the 443 documents, 185 of them have references in the QA dataset. We present the token distribution in Figure 6. Similarly, 70% of the documents could enter a context window of 500 tokens.

A.2 RoD-QA Statistics

The QA dataset totals 1,156 samples containing text and images modalities with and without legal

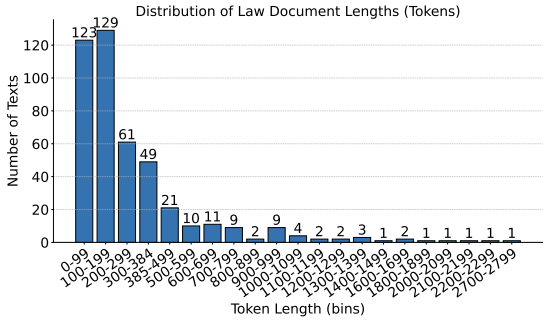


Figure 5: Number of tokens distribution of the entire RoD-Law corpus.

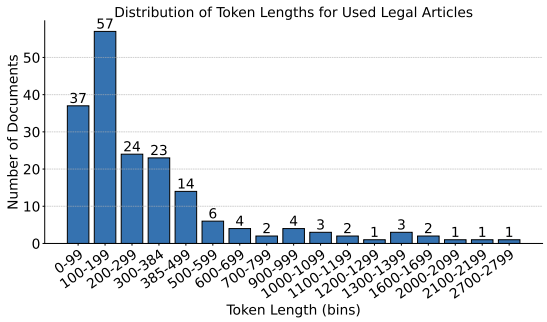


Figure 6: Number of tokens distribution on the documents referenced in RoD-QA.

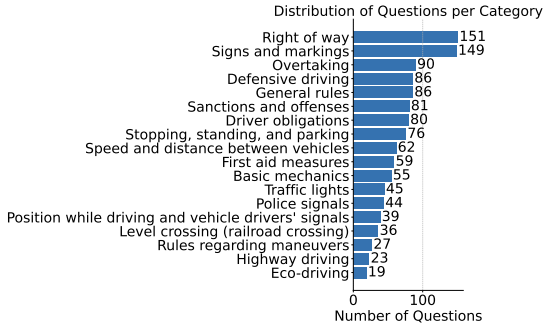


Figure 7: Number of questions per primary category

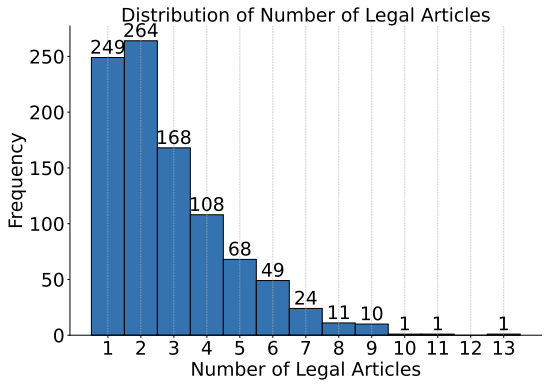


Figure 8: Distribution of legal article references per question

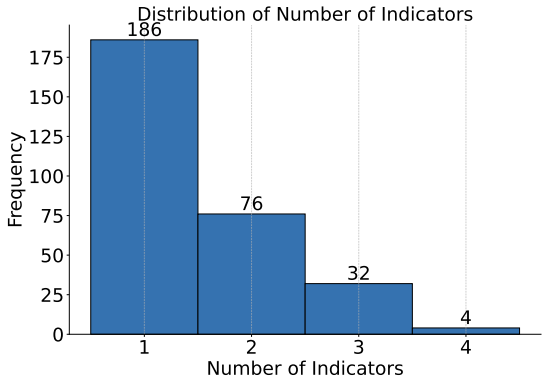


Figure 9: Distribution of traffic sign indicators per question

reference annotations, as presented in §3.2. In Figure 7, we showcase the distribution of questions per category. In total, there are 18 categories, most of them addressing *right of way* and *signs and markings*.

For the data annotated with legal references, we illustrate in Figure 8 the distribution per question. Most questions contain up to 10 references, which this also the main motivation to set $k = 10$ documents retrieved during experiments. In Figure 9, we present the distribution of the number of indicators illustrated in the questions containing images, where most questions show up to three indicators.

A.3 Skewed Distribution of Questions per Categories

We show the distribution of the legal documents per topic in Figure 10. Compared to Figure 7, where certain categories have fewer samples, such as *eco-driving*, *highway driving*, or *railroad crossing*, we want to find a similar mapping to the distribution of the laws. For example, *motorway driving* only has three legal articles associated with it. *Railway level crossing signage* has seven articles. Also, maneuvers in *rules regarding maneuvers* in Romanian legislation are a small set of actions (i.e., overtaking, stopping, stationary, parking, turning 180, and driving backwards). They involve only 19 different articles across all the questions, whereas we have 27.

We have also checked the coverage of the RoDTAL corpus (which includes all the existing articles in the useful law), and the 1,156 driving questions cover only 185 unique articles out of 443 (41%).

The official guidelines as to how the driving tests are created can be found in the MAI Order no. 268/2010 Article 10[6]. Paragraph (4) explicitly describes the content of the theoretical questions. The first points refer to general driving rules (for example, (a) states about legal provisions on road traffic such as road signs, priority rules, and speed restrictions), while the last points refer to vehicle safety equipment (point (h)) and environmental protection (point (i)). The ordering may also illustrate the importance.

Given this information, we conclude that:

- The underrepresented categories don't have a lot of associated laws to produce non-repetitive questions
- The driving tests reflect more daily situations and therefore will have more questions from daily situations.

A particular trait of our dataset is that it contains only the official driving test questions, as stated in the FAQ Q1[7] of *Scoala Rutiera*. Therefore, our question distributions are skewed due to (1) having only official questions and (2) having a focus on more daily situations rather than complete coverage of the law.

B Experimental Setup

B.1 Model Checkpoints

Table 10 presents the models and the checkpoints used during the experiments. We indicate the size as the number of parameters, context size in number of tokens, and the checkpoint on HuggingFace[8] or OpenAI Platform[9].

B.2 Hyperparameters

For the fine-tuning of the dense retriever model, we largely follow the settings from the original work, with several modifications specific to our setup. The model is trained using the Sentence Transformers library, with the following hyperparameters:

- **Number of epochs**: We train the first version of the model for 10 epochs, and the second one is early stopped after 1 epoch due to poor performance.

[6] https://legislatie.just.ro/Public/DetaliiDocument/124490
[7] https://www.scoalarutiera.ro/intrebari-frecvente.html
[8] https://huggingface.co/
[9] https://platform.openai.com

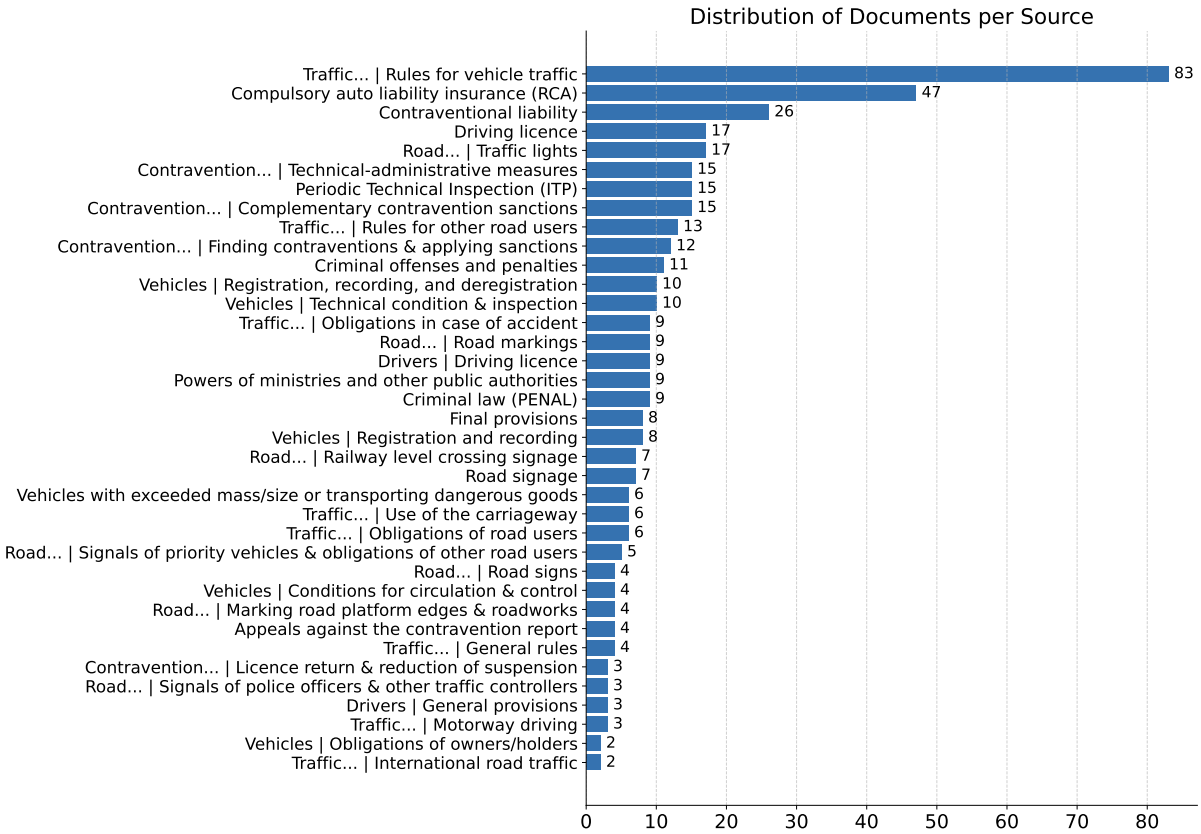


Figure 10: Number of legal documents per topic.

Model	Num. Params.	Ctx. Size	Checkpoint
$mE5_{small}$	118M	512	`multilingual-e5-small`
Passage Reranking Multilingual BERT	168M	512	`bert-multilingual-passage-reranking-msmarco`
Jina Reranker v2	278M	1024	`jina-reranker-v2-base-multilingual`
Mistral Small 3.1	24B	128k	`Mistral-Small-3.1-24B-Instruct-2503`
Gemma 3 27B Instruct	27B	128k	`gemma-3-27b-it`
GPT-4o mini	undisclosed	128k	gpt-4o-mini-2024-07-18
o4-mini (medium)	undisclosed	200k	o4-mini-2025-04-16

Table 10: Model checkpoints used during experiments.

- **Batch size**: Both the training and evaluation batch sizes are set to 64.
- **Learning rate schedule**: We use a warmup ratio of 0.1.
- **Mixed precision**: Training is performed with FP16 precision enabled, while BF16 is disabled.
- **Batch sampler**: We use `BatchSamplers.NO_DUPLICATES` to avoid duplicate samples within a batch, which is beneficial for in-batch negative sampling losses.
- **Loss function**: The MultipleNegativesRankingLoss (InfoNCE) is used, with $mE5_{small}$ as the base model.
- **Evaluation and saving**: The model is evaluated and checkpoints are saved every 100 steps, with only the best two checkpoints retained.
- **Early stopping**: The best model is loaded at the end of training based on the evaluation metric (`eval_cosine_recall@10`), with `greater_is_better=True`.
- **Random seed**: To ensure reproducibility, all relevant random seeds (`torch`, `numpy`, `random`, and `transformers`) are set to 42, and deterministic training options are enabled.

B.3 LLM Prompts

To maintain readability and keep the paper focused on the analysis and results, we make our LLM

prompts, both in Romanian and English, publicly available in the GitHub repository[10].

B.4 Hardware Infrastructure and Computational Costs

We train the open-source embedding model on Google Colab with one NVIDIA A100 GPU and do inference on the open-source LLM models on 4 NVIDIA A100 GPUs from the institutional cluster, using the *vllm* package for LLM inference and better scalability.

For the closed-source LLM experiments, the GPT-4o mini model is used with a seed set to 25 and a temperature of 0. The o4-mini model is used without the possibility to set temperature or seed, as for now. The total cost for LLM experiments using the OpenAI API is $115.72.

B.5 Information Retrieval Setting

Embedding model. We initially employ the $\text{mE5}_{\text{small}}$ multilingual model (Wang et al., 2024) to generate dense text embeddings for the legal corpus and user queries. This model supports Romanian and offers a sufficiently large context window to embed the relatively long articles typical of traffic law. It is also ranked among the top models on the MTEB leaderboard (Muennighoff et al., 2023) for retrieval tasks while supporting Romanian. We embed each article by concatenating the title metadata and content. We truncate the content if it is longer than the maximum context of 512, avoiding splitting into smaller chunks. During QA setups, where we use RAG, even if the retrieval section is done on the first 512 tokens, in the generation step we include the whole document. This decision does not affect generation in the end.

Query formatting. To improve retrieval performance, we experiment with various input formulations, including concatenating the question with its answer options, which led to better embedding-based similarity scores.

Reranking. Further improvements were sought by incorporating multilingual reranking models: Jina Reranker v2[11] and Passage Reranking Multilingual BERT (Nogueira and Cho, 2020).

Query rewriting. We also explore query rewriting using LLMs, specifically GPT-4o mini via API (OpenAI et al., 2024a), attempting to rephrase questions in a way that aligns with the embedding model's representation space.

Fine-tuning retriever. Following poor results from these techniques, we hypothesize that domain and language mismatch are a core bottleneck, specifically in the specialized and underrepresented Romanian legal language. To address this, we fine-tune the $\text{mE5}_{\text{small}}$ model on our dataset for 10 epochs, 80-20 split, using the InfoNCE loss (Oord et al., 2018), consistent with the model's original training regime. We constructed a training set of approximately 6,960 samples, comprising **positive pairs** (i.e., questions with their correct legal references), and **hard negatives** (i.e., derived from top candidates retrieved by the base model, but judged incorrect – positive aware hard-negative mining, 5 each). We also test with reranking on top of the fine-tuned retriever. This finetuning process was designed to fill the knowledge gap in the retriever. We believe that if certain articles become repealed in the future it won't impact retrieval – by simply removing it from the retrieval corpus. The retriever learns how to align driving questions and legal articles, not overfitting a lot, as shown in the results between Split 1 train & test.

Data augmentation. We further experiment with data augmentation via LLM-based synthesis. A few-shot prompt was created (2 examples of documents with associated questions and answers), and 1,000 sets of 2–6 legal documents were sampled. GPT-4o mini is used to generate 5 QA pairs per set. However, the LLM typically uses only 1–2 references per question, likely due to contextual incompatibility. After removing duplicates and entries with a similarity score over 0.98, we obtain 2,259 valid pairs, totaling 14,055 training samples. For this augmented dataset, we apply the same contrastive-based fine-tuning regime. The goal is not to leak any of the dataset distribution and statistics to our retriever and to validate based on the whole initial dataset.

B.6 QA and VQA Tasks Evaluation

In §4.5, we present that the exact match metric is mainly used for evaluating LLMs, similar to how a candidate is assessed during a driving test. The reason behind this is that during the exam, candidates aren't given partial scores for missing an option or selecting another one. This would also be critical to the driving capabilities, since giving the wrong answer would mean the candidate made a mistake and doesn't understand the law, which could lead to dangerous situations. Precision is

[10] https://anonymous.4open.science/r/rodtal-v2-F50C/

[11] Jina Reranker V2

paramount in this context.

However, in the actual context of LLMs, the *partially correct* answers are an important aspect in benchmarking the models (Kalai et al., 2025). Therefore, we also provide precision (P), recall (R), and F1 scores ($F1$) in Appendix E, which account for partially correct answers. Below, we describe how we compute these metrics.

For each question, we compute:

- TP – number of predicted answers that are actually correct
- FP – number of predicted answers that are not correct
- FN – number of correct answers that were missed

The per-question metrics are computed as follows:

$$P_{question} = \frac{TP}{TP + FP} \quad (1)$$

$$R_{question} = \frac{TP}{TP + FN} \quad (2)$$

$$F1_{question} = 2 \cdot \frac{P_{question} \cdot R_{question}}{P_{question} + R_{question}} \quad (3)$$

The overall metrics are then computed as the average over all questions, where N is the total number of questions:

$$P = \frac{1}{N} \sum_{i=1}^{N} P_{question_i} \quad (4)$$

$$R = \frac{1}{N} \sum_{i=1}^{N} R_{question_i} \quad (5)$$

$$F1 = \frac{1}{N} \sum_{i=1}^{N} F1_{question_i} \quad (6)$$

To illustrate how these metrics work, we provide some examples below. Suppose for a given question, the correct answers are $\{A, B\}$. Then, if:

- prediction $= \{A\} \rightarrow TP = 1, FP = 0, FN = 1 \rightarrow precision = 1.0, recall = 0.5, F1 = 0.667$ (partially correct)
- prediction $= \{A, C\} \rightarrow TP = 1, FP = 1, FN = 1 \rightarrow P = 0.5, R = 0.5, F1 = 0.5$ (partial with a penalty)
- prediction $= \{A, B, C\} \rightarrow TP = 2, FP = 1, FN = 0 \rightarrow P = 0.667, R = 1.0, F1 = 0.8$ (complete but with an extra wrong choice)
- prediction $= \{A, B\} \rightarrow TP = 2, FP = 0, FN = 0 \rightarrow P = 1.0, R = 1.0, F1 = 1.0$ (complete and correct answer)

C Discussions

C.1 Open-Source vs Closed-Source Models

Analyzing Table 6, if we compare experiments triplets (1,8,13), (2,9,14), (3,10,15), (4,11,16), or (5,12,17), the maximum performance gap for EM is 30% (experiment 3 and 10) and 65% on split 2. On split 2, this is primarily a failure of Mistral due to how we prompted the model. More specifically, split 2, compared to split 1, does not include legal references. However, our experiment still includes "`Here are the relevant documents [BLANK]`" (where `[BLANK]` represents that no document is passed). Due to this, Mistral would start and complete the missing documents instead of answering the questions without them.

This issue isn't seen in OpenAI's GPT-4o mini or Gemma 3 as they answer the questions, ignoring the missing documents. When comparing GPT-4o mini and Gemma 3 on the mentioned sets of experiments, their performance is relatively close, and in some cases Gemma 3 even outperforms. The real performance gap begins with OpenAI's o4-mini reasoning model, which indeed outperforms Gemma 3 (i.e., a non-reasoning model) compared to it. But this wouldn't be a fair comparison to do.

C.2 Question Difficulty

We discuss the question difficulty in our dataset from the perspective of LLM performance. We consider a question to be difficult in this setting if any model is unable to correctly predict it. For this, we take scenarios (6, 11, 12) in Table 6, and intersect the wrong questions from each model to find the ones that are always wrongly responded.

There is a total of 46 questions, with a higher value count of 7 and 4 in *general rules*, *Sanctions and offenses*, *right of way*, and *defensive driving*. Other than that, there is no observable pattern in these questions that would reflect that they are more difficult.

C.3 Answer Bias

Based on Table 11, the models don't show any bias regarding answer selection. The answer picking

Model	Split	Answer A		Answer B		Answer C	
		P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Gemma 3	1_test	33.68	33.56	36.32	36.91	30.00	29.53
Gemma 3	1_train	34.31	34.99	34.89	34.99	30.80	30.02
Gemma 3	2	34.78	33.17	40.10	41.21	25.12	25.63
Mistral	1_test	34.31	33.56	37.96	36.91	27.74	29.53
Mistral	1_train	35.22	34.99	32.59	34.99	32.19	30.02
Mistral	2	33.70	33.17	40.76	41.21	25.54	25.63
o4-mini	1_test	33.33	33.56	36.00	36.91	30.67	29.53
o4-mini	1_train	34.58	34.99	35.09	34.99	30.32	30.02
o4-mini	2	33.49	33.17	41.04	41.21	25.47	25.63

Table 11: Predicted (P) vs reference (R) distribution per model and split considering option letter frequency.

distribution matches the reference answers distribution. We can see that answer "C" typically has a lower occurrence as well in the references.

D Hallucination Analysis

D.1 Citations

We investigate how citations count in the model output correlate with the performance. For this analysis, we chose the scenario "Better prompt + RAG", which is where each model performs well. We extracted the citations using GPT 5-mini[12], a model different from the ones used during experiments, because the model tends to output citations in a different format than provided. We will discuss this in the following subsection. First, we looked at the number of citations per document, and we aggregated the results. We computed the average accuracy per group, and noticed a decrease in performance when models cite more unique documents; the findings are showcased in Figure 11. After this, based on the same grouping, we investigated how often these citations aren't part of the documents we provided in the retrieval step. If any response included a citation not given initially, we marked it as hallucinated. We then aggregated how many of the responses are hallucinated per the unique citations count, defining this as "Citation Hallucination Rate". Figure 11 shows a clear trend among all the models: citation hallucination rate increases constantly with the number of unique citations. OpenAI's o4-mini is having the most hallucination cases out of all the models, even if it retains the best performance.

Model	Split	Total	Rate (%)
Gemma	1_test	128	4.69
Gemma	1_train	510	2.35
Gemma	2	181	2.76
Mistral	1_test	128	25.78
Mistral	1_train	510	30.20
Mistral	2	181	28.18
o4-mini	1_test	128	2.34
o4-mini	1_train	510	1.76
o4-mini	2	181	0.55

Table 12: Instruction following in output formatting per model & split

D.2 Instruction Following

We analyze in Table 12 the model's instruction following consistency regarding the output format. We mark as incorrect if there is anything else present in the output after the "final response:" section other than the letters A, B and C. We can see the Gemma and o4-mini are consistent and follow instructions, while Mistral tends to have a higher incorrectness rate. Looking at outputs, for the 1_train split, out of 154 incorrect formatting, 115 of them are due to the model providing the answers at the beginning, and then providing the actual reasoning. In other examples, it adds even more explanations after giving the answers.

Another instruction-following criterion is that if the model cites articles using the given IDs we have. Looking at the outputs, instead of using the given ID of the law, it would usually rephrase it and cite it in a more official manner (e.g. "According to *article 4* from *OUG 195/2002*" instead of "According to *Regulation-4*" – the way it is cited is the wrong way as well, there is first an OUG (Emer-

[12] https://openai.com/index/introducing-gpt-5/

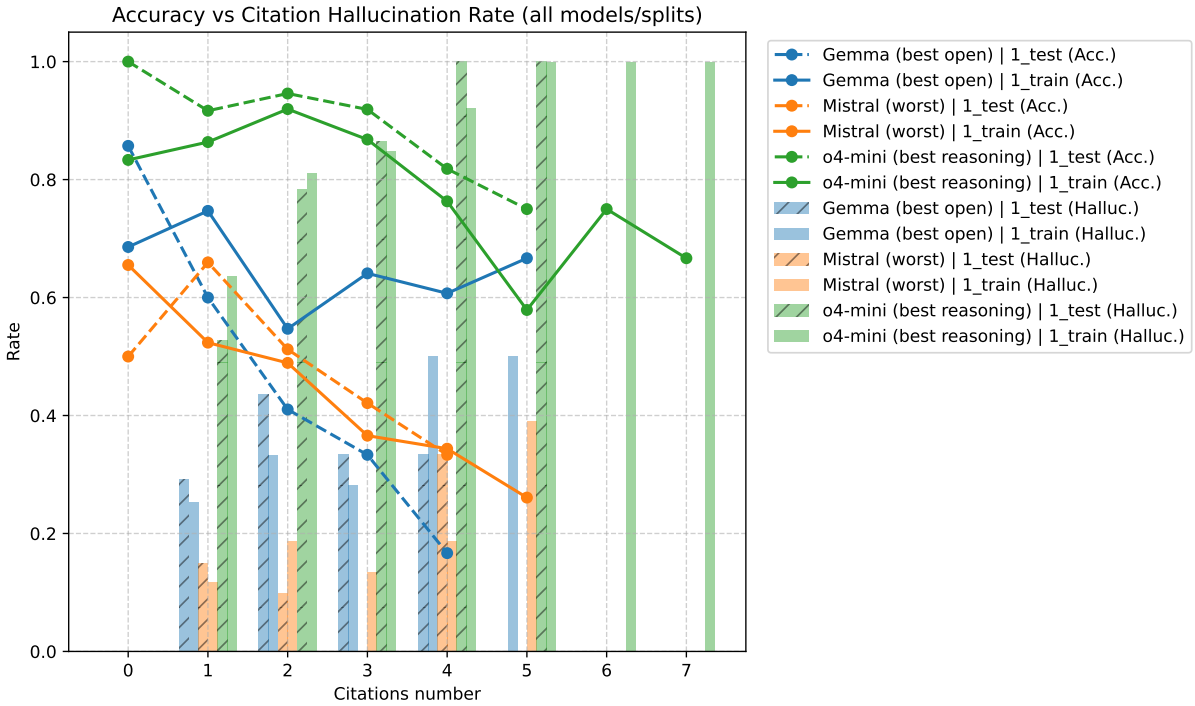


Figure 11: Accuracy vs Hallucination Rate on the number of unique citations present in output in the "Better Prompt + RAG" QA scenario.

Model	Split	Total	Rate (%)
Gemma 3	1_test	114	9.65
Gemma 3	1_train	475	6.53
Gemma 3	2	58	15.52
Mistral	1_test	118	2.54
Mistral	1_train	452	1.77
Mistral	2	77	1.30
o4-mini	1_test	125	11.20
o4-mini	1_train	498	10.44
o4-mini	2	80	8.75

Table 13: Wrong citation format per model and split

gency Ordinance) and then it has Regulations on how to apply it, therefore citing the Regulations as the OUG is wrong in the first place).

In Table 13 we analyzed in how many of the cases which contain citation, there is no occurrence of the correct pattern. For e.g. a model can cite "According to *article 4* from *OUG 195/2002* (*Regulation-4*)", so including the correct format along the incorrect one will be counted as correct. Mistral is the model which follows citation format best, while Gemma and o4-mini in avg 10% of the cases, they don't follow format. We can say, based on this, that the model tends to follow the training data format, instead of the given format, for regulatory data.

D.3 LLM-as-a-Judge for Hallucination Detection

We use the same OpenAI o4-mini, Gemma 3 27B Instruct and Mistral Small 3.1 Instruct to perform an LLM-as-a-Judge (Zheng et al., 2023) evaluation regarding hallucinations. For the samples, we used the outputs for scenarios (6, 11, 12) in 6, and we filter for those which were answered wrong, leaving us with 789 samples to analyze.

We identify a few criteria that can be judged on a True or False basis.

For this step, we use two methodologies. First, we examine the outputs to identify some common mistakes. We propose 2 initial criteria which stood out for us and will be denoted in what follows as **C4** and **C5**.

Our other methodology was to first use an LLM as a judge (Gemma 3) to identify (i.e., not classify) the mistakes in the initial outputs, in the format of a few bullet points, getting 3-4 mistakes per output. We then split the text based on the bullet points and newlines, embed them using $mE5_{small}$, and then use **KMeans** (Lloyd, 1982) to identify 15 clusters with random choice. We examine these clusters and choose three that seem more consistent.

Based on both methodologies, we get 5 final

criteria which we want to analyze:

- C1: **Over-interpretation.** The model introduces legal rules, conditions, or consequences that are not explicitly stated in the normative texts.

- C2: **Negligence.** The model ignores relevant legal information, such as exceptions, special conditions, or complementary provisions.

- C3: **Erroneous interpretation.** The model misreads or misapplies existing legal rules without inventing new ones.

- C4: **Values.** The model applies a wrong legal threshold, numerical value, or standard, or uses the correct value in the wrong context.

- C5: **Recommendations** The model confuses the legal force of recommendations, treating them as strict obligations or vice versa.

We can see in Table 14 the agreement for each criteria from the judges. C3 is the best criteria we have, followed by C2 and C1. These have a strong majority agreement on a True basis.

C4 and C5 have a very strong agreement on the False basis, although C4 still has a decent agreement on True.

Looking at the Pairwise Correlations and Cohen's Kappa (Kohen, 1960) in Tables 15 and 16, there is a strong correlation on criteria C4 and C5 for all the judges, which also reflect on the False agreement previously. C3 is next in terms of correlation, while C1 and C2 have less of a correlation.

We can observe the distribution in Figures 12 and 13 for the samples where the LLMs agree on hallucinations based on their question category. For C1, C2 and C3, we can see a higher ratio for "Driver obligations", "Railroad crossing", "General rules" and "Overtaking". For C4, we see a higher ratio on "Stopping, standing, and parking", "Sanctions and offenses", "Rules regarding maneuvers" and "Overtaking". For C5 we see the ratio is dominant for "Defensive driving" and "Overtaking", but this category also had the least agreement and samples. These samples, based on the previous filtering, are also all marked wrong during QA evaluation.

We added a few examples where all three models had agreement for one of the criteria in Appendix F.

D.4 Captioning Analysis

We decide to inspect the generated captions based on the following criteria

- Is the caption wrong (i.e., even partially?) – it contains a wrong description, added elements, or misses something very important

- Does the caption contain any legal interference – the caption also contains legal obligations, such as "therefore the driver has to ..."

- Is the legal interference wrong – if the previous option is True, then is this interference correct or not

We aim to analyze why VQA performance is lower when a caption is also present in the prompt, along with the image.

We identify and perform analysis based on five categories:

- (1) Correct caption, no legal interference and therefore no wrong legal interference

- (2) Correct caption, it has legal interference but there is no wrong legal interference

- (3) Correct caption, it has legal interference and there is wrong legal interference

- (4) Wrong caption, no legal interference and therefore no wrong legal interference

- (5) Wrong caption, it has legal interference and there is wrong legal interference

Based on Table 17, and on the categories (1) and (2), where the caption is correct and if it has legal interference, that is also correct, we can see there is not a high difference between the experiments (3-7%). In the other case, (3) where the caption is correct but the legal interference is wrong, the performance varies by 9-12%. Going further, with categories (4) and (5), where the captions is wrong descriptive-wise and also may or may not contain wrong legal interference, the performance varies between 14-27% when looking at the experiments with and without the added captions.

With this analysis in mind, the performance degradation that comes from using captions is caused from the captions being either wrong and/or containing wrong legal interference as well.

Criteria	Agreement (%)	True (%)	Majority True (%)	False (%)	Majority False (%)
C1	28.0	27.3	69.8	0.3	41.8
C2	23.2	20.7	70.5	3.2	39.2
C3	43.1	41.0	85.6	4.4	27.7
C4	85.0	20.9	54.9	84.1	93.3
C5	90.9	4.0	30.7	90.3	97.5

Table 14: LLM-as-a-Judge Agreement

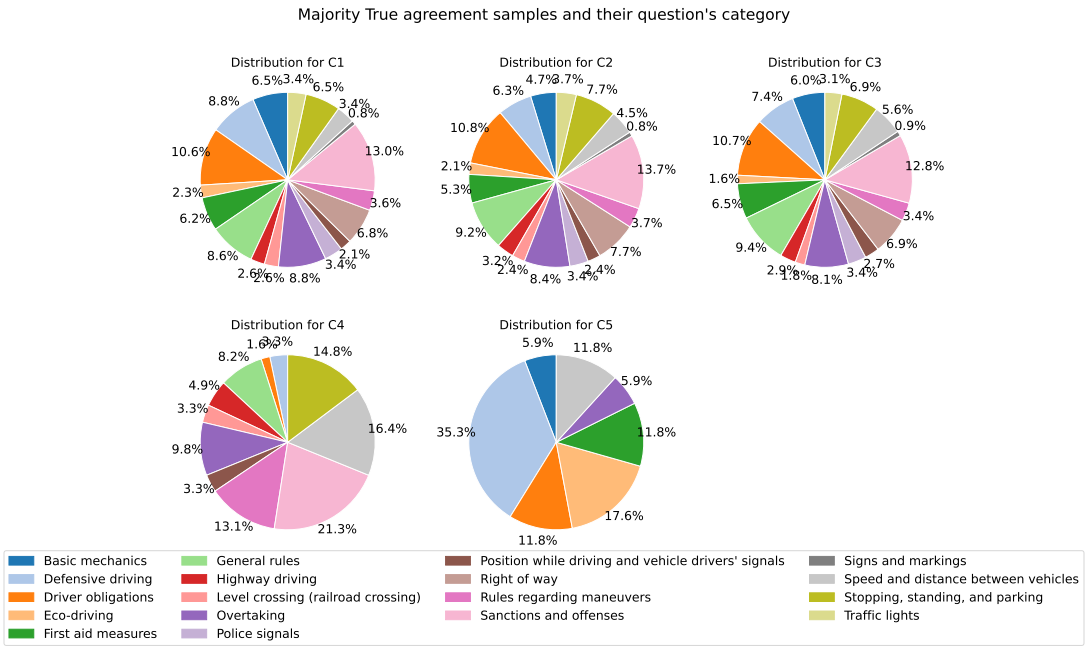


Figure 12: Majority True agreement samples and their question's category

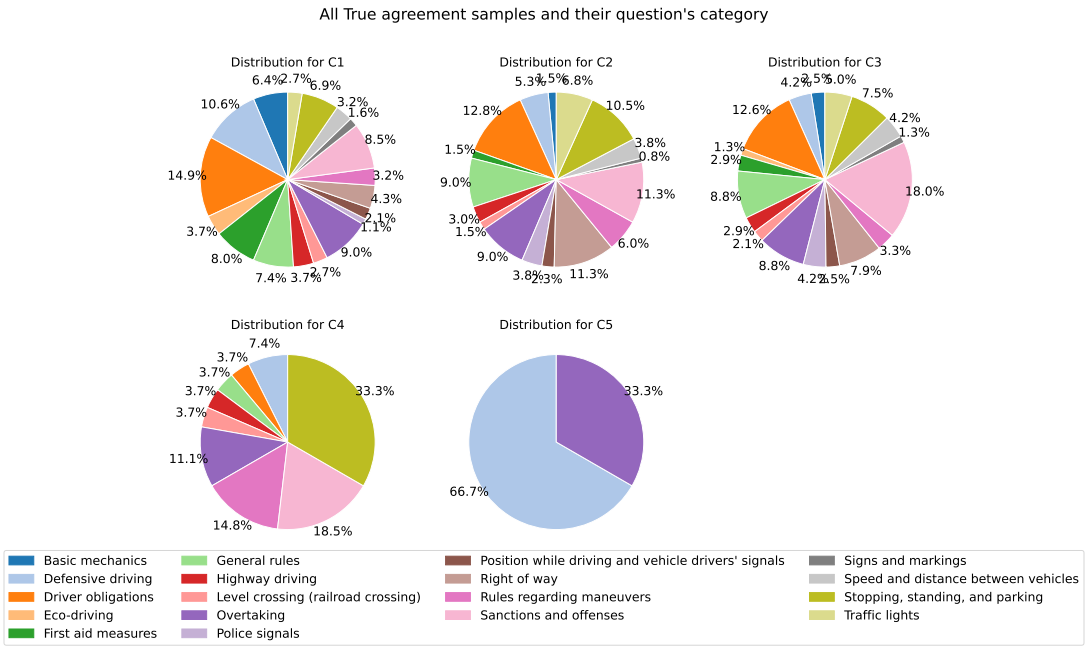


Figure 13: All True agreement samples and their question's category

We can also see in the same table, when the LLM is given the caption (e.g. comparing image and caption versus image only) the model will have less reasoning steps on average. This means that

Criteria	G-O	M-G	M-O
C1	3.85	6.18	22.09
C2	2.97	12.78	17.62
C3	15.54	19.24	22.86
C4	49.99	59.38	49.46
C5	20.97	42.42	30.81

Table 15: LLM-as-a-Judge Pairwise Correlations (M - Mistral, G - Gemma, O - OpenAI)

Criteria	G-O	M-G	M-O
C1	0.30	0.76	19.83
C2	0.79	7.26	12.92
C3	8.36	19.12	13.59
C4	46.09	58.87	43.10
C5	17.02	37.46	17.34

Table 16: LLM-as-a-Judge Pairwise Cohen's Kappa (M - Mistral, G - Gemma, O - OpenAI)

by providing a caption, the model will attend less the image and will reason less over the question, producing less strong results.

Category	Scenario	Total	Accuracy (%)	Reasoning count
(1)	C	41	80.4	1.80
(1)	C and I	41	82.9	1.56
(1)	I	41	87.8	1.58
(2)	C	167	83.8	1.26
(2)	C and I	167	83.8	1.29
(2)	I	167	86.8	1.70
(3)	C	33	63.6	1.48
(3)	C and I	33	66.6	1.63
(3)	I	33	75.7	2.00
(4)	C	30	46.6	2.30
(4)	C and I	30	56.6	2.33
(4)	I	30	73.3	2.26
(5)	C	116	50.8	1.91
(5)	C and I	116	52.5	1.87
(5)	I	116	64.6	2.13

Table 17: Caption category accuracy and reasoning count per scenario (C - caption, I - image)

E Additional Experimental Results

Method	Metric	Split 1		Split 2
		Train	Test	Test
(1) GPT-4o mini + CoT + RAG	Precision	77.0	74.7	85.5
(2) GPT-4o mini + CoT w/o RAG	Precision	69.0	67.3	85.2
(3) GPT-4o mini + CoT + Ideal RAG	Precision	76.5	76.8	82.7
(4) GPT-4o mini + CoT + RAG + better prompt	Precision	80.9	85.8	87.7
(5) GPT-4o mini + RAG + better prompt w/o CoT	Precision	68.5	70.2	79.2
(6) o4-mini + CoT + RAG + better prompt	Precision	**91.7**	**95.3**	**91.2**
(7) o4-mini + CoT + better prompt w/o RAG	Precision	76.8	77.1	89.2
(8) Mistral + CoT + RAG	Precision	65.2	66.8	70.2
(9) Mistral + CoT w/o RAG	Precision	80.3	78.6	91.0
(10) Mistral + CoT + Ideal RAG	Precision	46.9	51.7	18.5
(11) Mistral + CoT + RAG + better prompt	Precision	65.5	69.5	66.6
(12) Mistral + RAG + better prompt w/o CoT	Precision	88.6	86.8	93.4
(13) Gemma 3 + CoT + RAG	Precision	76.0	68.5	83.9
(14) Gemma 3 + CoT w/o RAG	Precision	63.0	58.4	75.5
(15) Gemma 3 + CoT + Ideal RAG	Precision	75.3	71.5	84.6
(16) Gemma 3 + CoT + RAG + better prompt	Precision	80.0	70.8	87.5
(17) Gemma 3 + RAG + better prompt w/o CoT	Precision	73.3	67.6	84.3
(1) GPT-4o mini + CoT + RAG	Recall	92.8	92.7	96.9
(2) GPT-4o mini + CoT w/o RAG	Recall	87.0	90.2	97.3
(3) GPT-4o mini + CoT + Ideal RAG	Recall	91.7	94.9	94.6
(4) GPT-4o mini + CoT + RAG + better prompt	Recall	90.7	92.8	94.6
(5) GPT-4o mini + RAG + better prompt w/o CoT	Recall	93.6	92.6	95.5
(6) o4-mini + CoT + RAG + better prompt	Recall	93.0	97.0	95.1
(7) o4-mini + CoT + better prompt w/o RAG	Recall	82.7	84.9	92.4
(8) Mistral + CoT + RAG	Recall	**97.6**	97.5	**98.6**
(9) Mistral + CoT w/o RAG	Recall	96.7	95.7	**98.6**
(10) Mistral + CoT + Ideal RAG	Recall	94.4	95.0	84.9
(11) Mistral + CoT + RAG + better prompt	Recall	97.1	**97.9**	97.2
(12) Mistral + RAG + better prompt w/o CoT	Recall	97.3	97.2	97.9
(13) Gemma 3 + CoT + RAG	Recall	89.4	87.1	93.9
(14) Gemma 3 + CoT w/o RAG	Recall	76.0	79.3	84.1
(15) Gemma 3 + CoT + Ideal RAG	Recall	88.8	88.9	94.6
(16) Gemma 3 + CoT + RAG + better prompt	Recall	90.5	87.2	90.7
(17) Gemma 3 + RAG + better prompt w/o CoT	Recall	88.8	88.3	92.0
(1) GPT-4o mini + CoT + RAG	F1-score	81.9	80.1	85.5
(2) GPT-4o mini + CoT w/o RAG	F1-score	74.1	74.1	89.1
(3) GPT-4o mini + CoT + Ideal RAG	F1-score	81.2	83.7	86.4
(4) GPT-4o mini + CoT + RAG + better prompt	F1-score	83.6	87.4	89.8
(5) GPT-4o mini + RAG + better prompt w/o CoT	F1-score	76.4	76.6	84.1
(6) o4-mini + CoT + RAG + better prompt	F1-score	**91.7**	**95.7**	92.2
(7) o4-mini + CoT + better prompt w/o RAG	F1-score	78.1	78.7	89.9
(8) Mistral + CoT + RAG	F1-score	69.1	70.9	73.6
(9) Mistral + CoT w/o RAG	F1-score	84.8	83.3	93.8
(10) Mistral + CoT + Ideal RAG	F1-score	50.8	55.2	23.4
(11) Mistral + CoT + RAG + better prompt	F1-score	68.2	72.4	69.6
(12) Mistral + RAG + better prompt w/o CoT	F1-score	92.0	90.7	**94.5**
(13) Gemma 3 + CoT + RAG	F1-score	79.7	73.8	86.7
(14) Gemma 3 + CoT w/o RAG	F1-score	66.5	64.2	77.9
(15) Gemma 3 + CoT + Ideal RAG	F1-score	79.2	76.3	87.5
(16) Gemma 3 + CoT + RAG + better prompt	F1-score	82.9	75.1	88.2
(17) Gemma 3 + RAG + better prompt w/o CoT	F1-score	77.8	73.3	86.3

Table 18: Precision, recall, and F1-score on the IR/QA RAG pipeline.

Method	Metric	Split 3	Split 4
(1) o4-mini + Caption + QA + CoT w/o RAG	Precision	65.8	72.5
(2) o4-mini + Image + QA + CoT w/o RAG	Precision	72.4	76.8
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	Precision	66.3	74.6
(4) o4-mini + Caption + QA + CoT + Ideal RAG	Precision	70.7	80.3
(5) o4-mini + Image + QA + CoT + Ideal RAG	Precision	78.6	79.6
(6) o4-mini + Image + Caption + QA + Ideal RAG	Precision	73.1	78.9
(7) o4-mini + Caption + QA + CoT + RAG	Precision	68.5	76.1
(8) o4-mini + Image + QA + CoT + RAG	Precision	**76.3**	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	Precision	69.8	78.2
(1) o4-mini + Caption + QA + CoT w/o RAG	Recall	67.4	73.2
(2) o4-mini + Image + QA + CoT w/o RAG	Recall	73.4	78.9
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	Recall	67.7	74.6
(4) o4-mini + Caption + QA + CoT + Ideal RAG	Recall	71.5	81.7
(5) o4-mini + Image + QA + CoT + Ideal RAG	Recall	**79.4**	80.3
(6) o4-mini + Image + Caption + QA + RAG	Recall	74.4	78.9
(7) o4-mini + Caption + QA + CoT + RAG	Recall	69.6	76.1
(8) o4-mini + Image + QA + CoT + RAG	Recall	76.9	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	Recall	70.3	78.9
(1) o4-mini + Caption + QA + CoT w/o RAG	F1-score	66.4	72.8
(2) o4-mini + Image + QA + CoT w/o RAG	F1-score	72.7	77.5
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	F1-score	66.8	74.6
(4) o4-mini + Caption + QA + CoT + Idea RAG	F1-score	71.0	80.8
(5) o4-mini + Image + QA + CoT + Ideal RAG	F1-score	**78.9**	79.8
(6) o4-mini + Image + Caption + QA + Ideal RAG	F1-score	73.5	78.9
(7) o4-mini + Caption + QA + CoT + RAG	F1-score	68.9	76.1
(8) o4-mini + Image + QA + CoT + RAG	F1-score	76.5	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	F1-score	69.9	78.4

Table 19: Precision, recall, and F1-score on the VIR/VQA RAG pipeline using strategy (5) for laws and strategy (2*) for indicators.

F Question and Answer Samples

This section highlights key examples of observed failure modes in strategy 1 (base LLM output) and their successful mitigation by strategy 4. Each example demonstrates a specific challenge in legal question-answering for LLMs and how a refined approach (strategy 4) can address these issues. Check the comments for every question for our observations.

Difficult Questions: These examples illustrate strategy 1's difficulty in discerning subtle nuances or misleading elements within questions or options, leading to incorrect or incomplete answers. Strategy 4 consistently provides more precise and legally accurate responses.

We make these samples prompts publicly available in this repository [13].

LLM as a judge outputs

- **Gemma / 0692408a**:
  - **Question:** What should you consider when intending to overtake both cyclists and groups of pedestrians moving on the roadway?
  - **Options:**
    - A. You may overtake only after pedestrians and cyclists signal you to do so.
    - B (True). You must comply with the legal provisions.
    - C. Cyclists and pedestrians may only be overtaken if the lateral distance is 3 meters.
  - **Model answer:** Correct answers: B, C.
    B is correct because overtaking cyclists and pedestrians must comply with the general legal provisions regarding overtaking (OUG-46).
    C is considered correct because, although not explicitly stated for pedestrians, the lateral distance of 3 meters is mandatory when overtaking cyclists according to the Traffic Regulations.
    A is incorrect because it is not necessary for pedestrians or cyclists to signal before being overtaken; it is the driver's responsibility to ensure safe overtaking.
  - **Hallucination criteria:** Over-interpretation.
  - **Gemma judge:** The answer claims that a fixed lateral distance of 3 m is mandatory when overtaking cyclists, based on the Traffic Regulations. However, the regulations do not set a fixed value; they only mention maintaining a "sufficient lateral distance" (Regulation 118 c), with additional rules depending on speed (under/over 50 km/h) regarding crossing the lane marking.

[13] https://anonymous.4open.science/r/rodtal-v2-F50C/

- **Mistral judge:** The person generalized the 3 m lateral distance rule (applicable to cyclists) and applied it to pedestrians, although Regulation 118 c) does not specify such a distance for pedestrians.
- **o4-mini judge:** The candidate incorrectly stated that the lateral distance when overtaking cyclists is 3 m, while Regulation 118 c) only requires a "sufficient lateral distance" without specifying a fixed number.

- **o4-mini / cef83a0f**:
  - **Question:** You are required to use dipped headlights during the daytime:
  - **Options:**
    - A (True). On all categories of public roads.
    - B. Only on county, communal, and local roads.
    - C. Only on motorways, expressways, and European national roads.
  - **Model answer:** The model reasoned that Article 32(2) of OUG 44/2002 establishes that the obligation to use dipped headlights unconditionally during the day exists only on motorways, expressways, and European national roads. Elsewhere, the use of dipped headlights depends on lighting conditions (Regulation 114/2002). The model selected option C as correct.
  - **Hallucination criteria:** Negligence.
  - **Gemma judge:** The answer misinterprets Article 32(2) of OUG 44/2002, assuming it establishes the only obligation regarding daytime dipped headlights, whereas it applies only to motor vehicles on motorways, expressways, and European national roads.
  - **Mistral judge:** The person neglected Article 32(3) of OUG 44/2002, which requires motorcycles and mopeds to use dipped headlights on all public roads, regardless of road category.
  - **o4-mini judge:** Article 32(3) of OUG 44/2002 was not considered; it mandates that motorcycles and mopeds must use dipped headlights on all public roads.

- **o4-mini / ee8de379**:
  - **Question:** What does a continuous longitudinal line separating traffic directions mean?
  - **Options:**
    - A. It guides drivers in low-visibility conditions.
    - B (True). It prohibits crossing it.
    - C. It prohibits overtaking on the section where it is applied.
  - **Model answer:** The model argued that:
    – Article 19(2) of Regulation 77 states that a continuous line prohibits crossing.
    – Article 103(i) of Regulation 120 prohibits overtaking when crossing the continuous line.
    Therefore, it concluded both B and C are correct.
  - **Hallucination criteria:** Erroneous interpretation.
  - **Gemma judge:** The person incorrectly equated "prohibiting crossing" with "prohibiting overtaking" entirely, without considering overtaking that does not cross the line.
  - **Mistral judge:** The person interpreted the continuous line as an absolute ban on overtaking, ignoring Regulation 120, which allows overtaking if the line is not crossed.
  - **o4-mini judge:** The candidate misapplied Article 120(i), interpreting the continuous line as prohibiting any overtaking, although the text only forbids overtaking when the line is crossed.

- **Gemma / 2d7567f7**:
  - **Question:** U-turns are prohibited:
  - **Options:**
    - A (True). On pedestrian crossings.
    - B. Less than 50 m from pedestrian crossings.
    - C. At intersections, including roundabouts.
  - **Model answer:** According to the Traffic Regulations, U-turns are prohibited:
    - on pedestrian crossings (letter e),
    - in intersections, including roundabouts (letter b).

    Less than 50 m from pedestrian crossings applies to stopping, not U-turns.
    Correct answers: A, C.
  - **Hallucination criteria:** Values
  - **Gemma judge:** The answer confused the 50 m stopping distance rule (Regulation 142 b) with U-turn prohibitions.
  - **Mistral judge:** Regulation 142(e) prohibits stopping less than 25 m before and after pedestrian crossings, but no evidence supports a 50 m U-turn prohibition near crossings.
  - **o4-mini judge:** The answer confused the 50 m stopping rule with U-turn rules; 50 m applies to stopping, not U-turns.

- **Mistral / 3d4b79a3**:
  - **Question:** What should you consider when approaching a school and noticing that students are leaving the building?
  - **Options:**
    - A. The law provides no obligations.
    - B (True). Students may be inattentive and suddenly appear in front of the car.
    - C. Students will only cross at pedestrian crossings.
  - **Model answer:** The model answered B and C, reasoning that:
    - Students, especially younger ones, can be inattentive and may suddenly step onto the road.

    - By law, pedestrians must cross at designated pedestrian crossings; however, drivers should not rely exclusively on this rule.
    - It stated that the law does not impose specific obligations but recommended increased caution and reduced speed.
  - **Hallucination criteria:** Recommendations.
  - **Gemma judge:** The answer described reducing speed and paying attention as merely "recommended", downplaying the legal obligation under Regulation 123(h).
  - **Mistral judge:** The person treated a legal obligation (Regulation 123 h) to reduce speed in school zones) as just a recommendation.
  - **o4-mini judge:** Reducing speed in school zones was presented as a recommendation, but Regulation 123(g) makes it a legal obligation.

G Detailed Error Analysis

G.1 Information Retrieval

Looking more in depth, if we compare strategies (1) and (3) in detail per category (see Figures 14 and 15), we can see the strengths of the first strategy. In some categories, while recall went higher on the training split, as these experiments don't involve training, the score went down on the test split. This could also be due to their small size, and having it split across so many categories decreases the number of samples per category even more.

On (6) from Figure 16, where we employ fine-tuning, we can see an overall improvement in all categories. We observe that categories such as *Highway Driving* and *Defensive Driving* underperform on the test set, due to a smaller share of the dataset, which would make it hard to learn relevant features. Similarly, the test set score is also smaller in *Sanctions and Offenses* and *General Rules*. A fully comparative analysis between all the strategies can be seen in Figures 18 and 19.

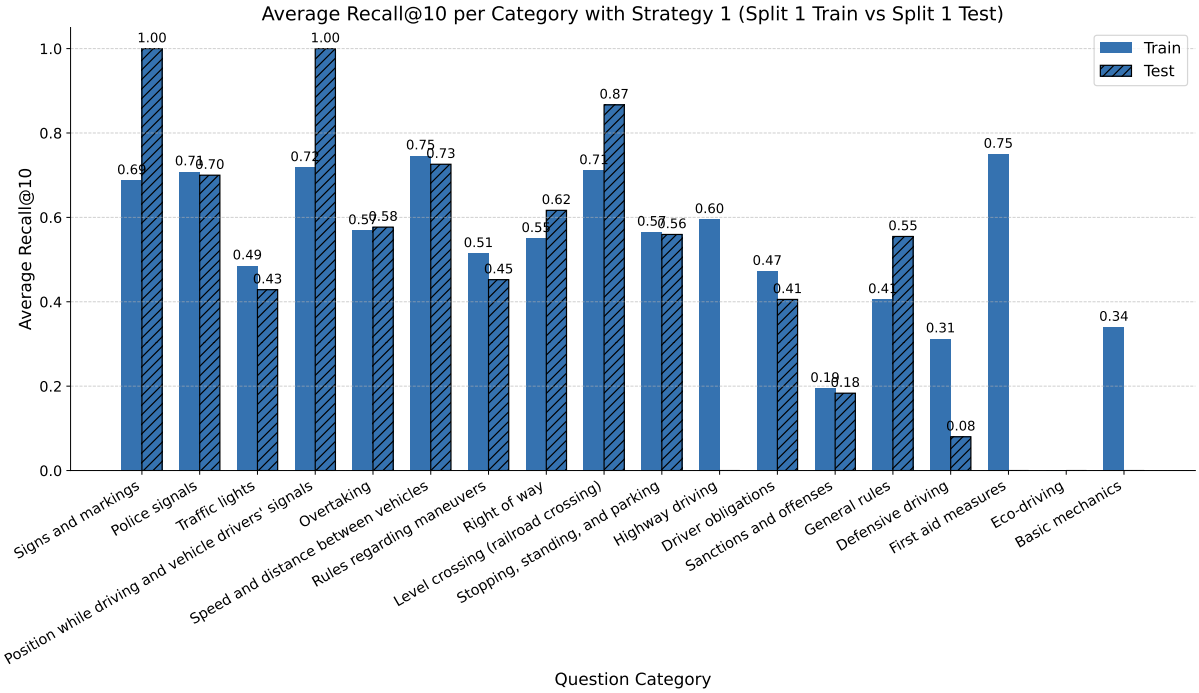


Figure 14: Recall@10 Analysis for IR task with strategy 1

G.2 Question Answering

Looking at the performances per category, in Figures 20, 21 and 22, we can see that in most categories the RAG ablation has a significant impact or in some cases similar performances. However, in "Eco Driving" and "Basic Mechanics", categories which don't need legal groundings but rather general other kinds of knowledge, the RAG is downgrading the performance (due to context bloat). But there a not a lot of entries in these categories in the Split 1 train and test. In Split 2, where we have more entries, we have similar or closer results between the strategies. We can also see that the Ideal RAG compared to our

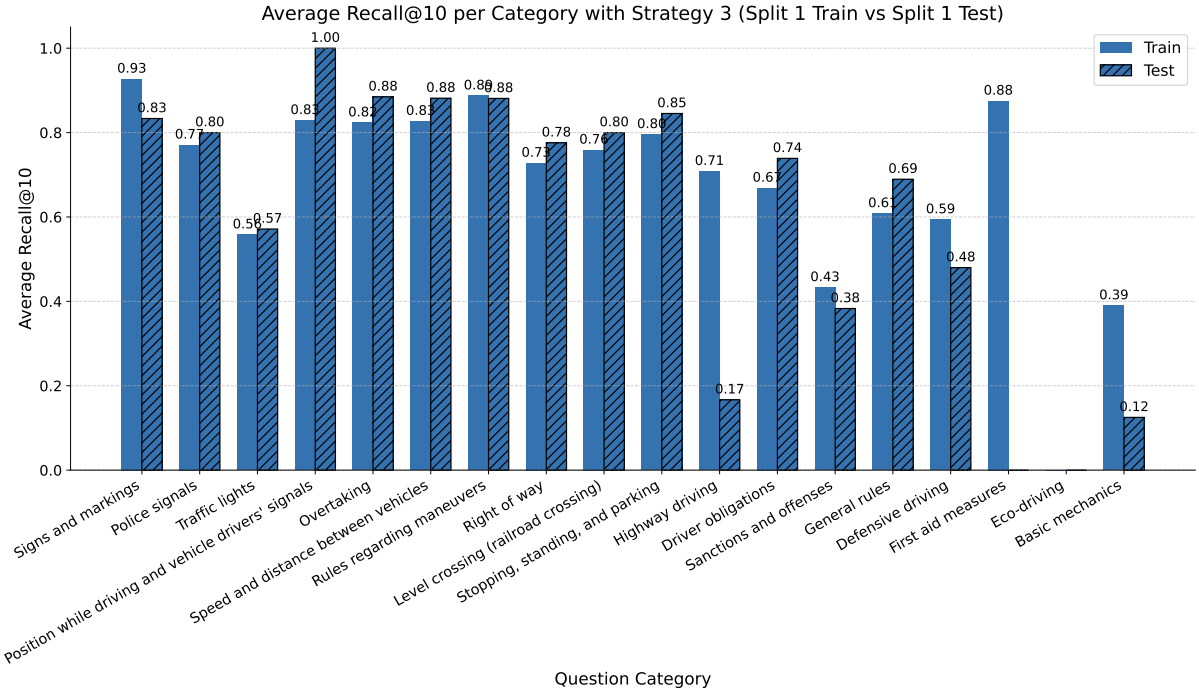


Figure 15: Recall@10 Analysis for IR task with strategy 3

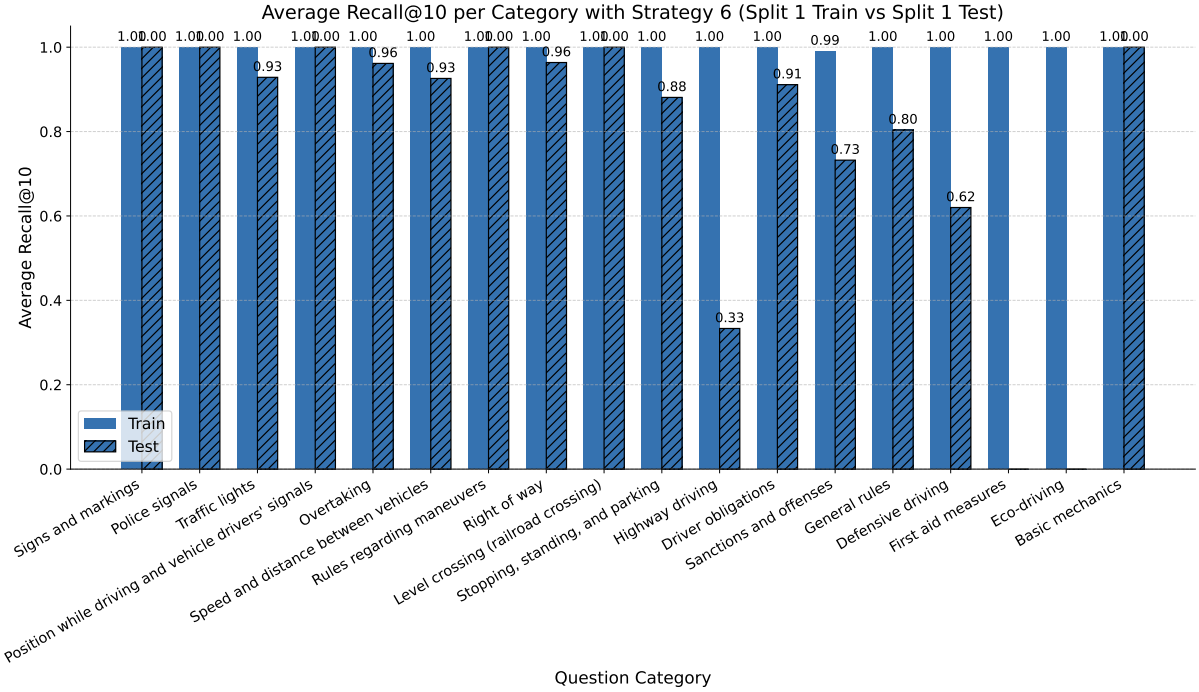


Figure 16: Recall@10 Analysis for IR task with strategy 6

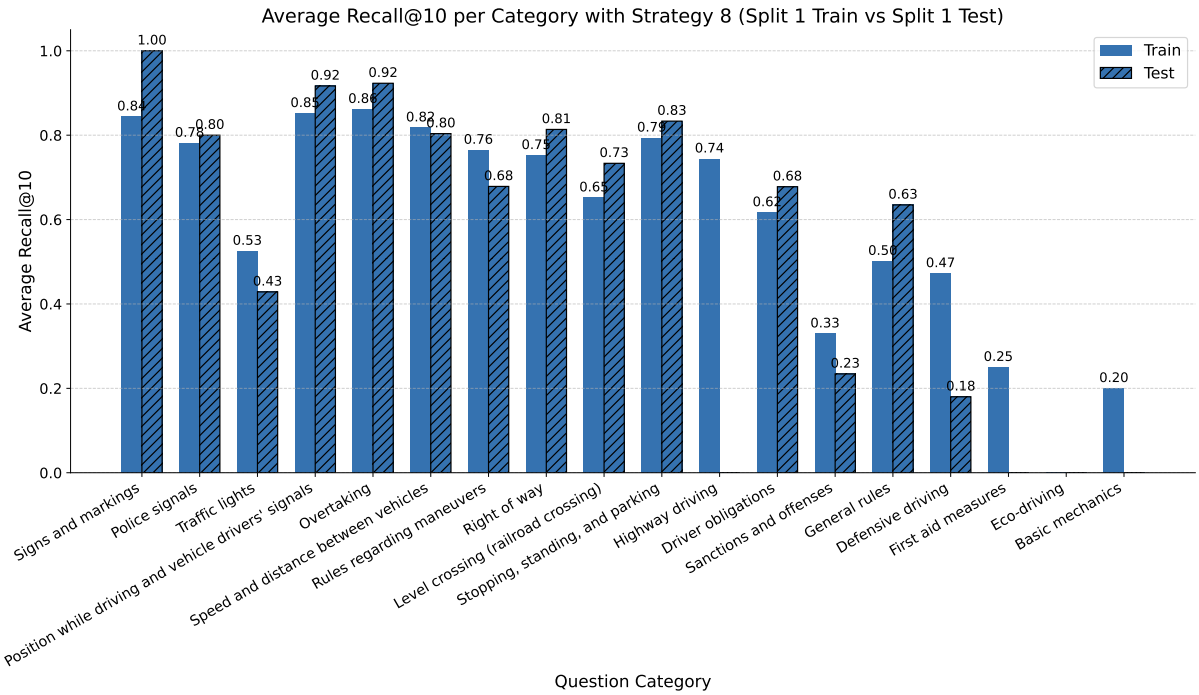


Figure 17: Recall@10 Analysis for IR task with strategy 8

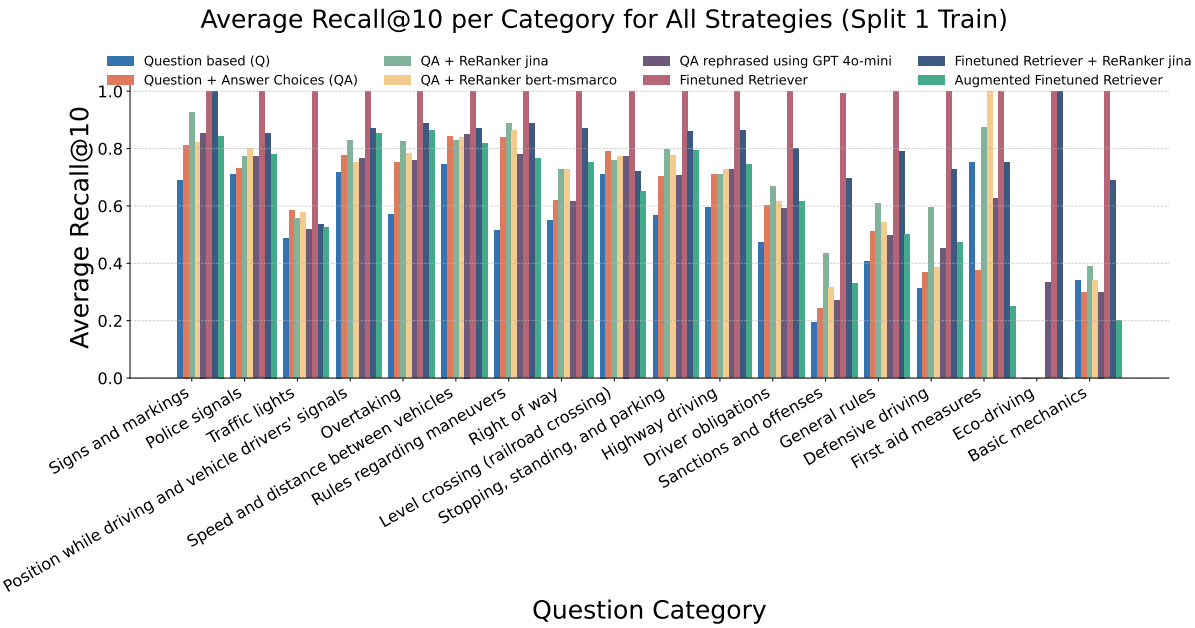


Figure 18: Recall@10 Analysis for IR task with all strategies on Split 1 Train

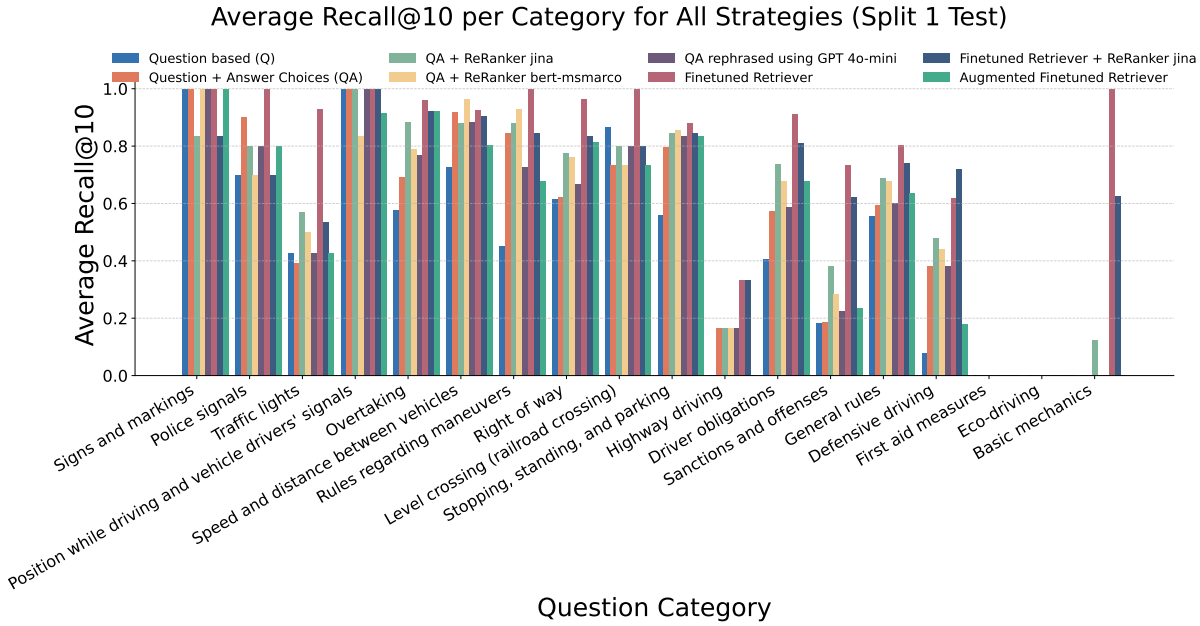


Figure 19: Recall@10 Analysis for IR task with all strategies on Split 1 Test

proposed solution isn't making a significant impact, showing that better retrieval doesn't lead to better QA in our case. We can see that the reasoning models perform better than normal ones, and the RAG is still making an impact in most of the categories.

For the Mistral model, we can see in Figures 23, 24, and 25 that not employing RAG helps in most of the categories. This is likely a flaw of the Mistral model, as its outputs also tend to respond in the wrong format or fail to respect instructions, instead continuing to write article paragraphs rather than responding to the query.

If we look at the number of selected answers, in Figures 29, 30 and 31, most of the time the models selects either the right amount of answers or over the amount of answers, but in very few cases less than the amount. Similarly to the recall vs precision comparison, these experiments validate that the model mostly chooses the right answers but sometimes picks more than it should when it makes errors.

If we look at the number of reasoning steps, in Figures 32, 33 and 34, when we don't include RAG, the model needs more steps to arrive at a final answer. This seems like a normal behavior, because in the RAG case it gets missing information instead of reasoning about it. We can see it needs most of the reasoning steps in eco-driving or first-aid, where law documents don't really help.

G.3 Visual Information Retrieval

Per category analysis of retrieved laws, in Figure 37, we can see it performs worse on *signs and markings*, *general rules*, and *defense driving*. The first one is more traffic sign-intensive, so it implies that the model focuses more on the signs instead of the laws related to the question.

G.4 Visual Question Answering

Looking at Figures 38 and 39, we see the model has lower performance in *Position while driving as vehicle drivers' signals* and *Stopping, standing and parking*. If we look at secondary categories, in Figures 40 and 41, we can see worse performance in the *Aerial* and similar performance in the other two. Similarly, this is the category where, in Figures 46 and 47, the models tend to use more reasoning steps than the other. If we look at the number of selected answers, in Figures 42 and 43, the model tends not to select more than enough answers (even if incorrect). We can see the comparison of reasoning steps over the primary category in Figures 44 and 45.

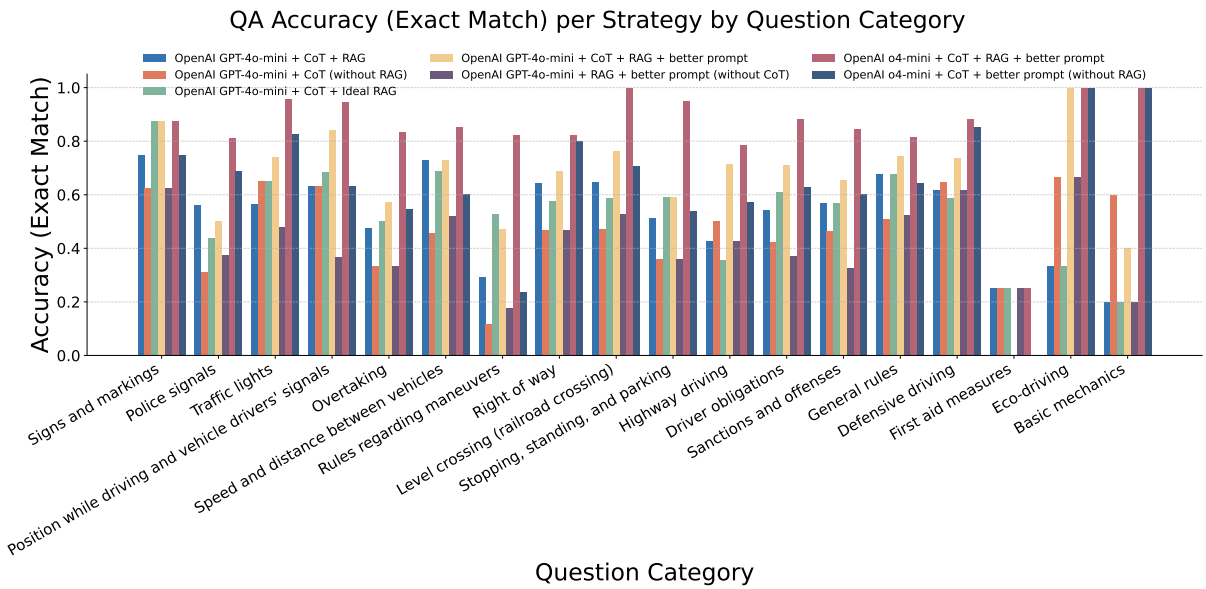


Figure 20: OpenAI models Exact Match score for QA per Strategy and Category on Split 1 Train

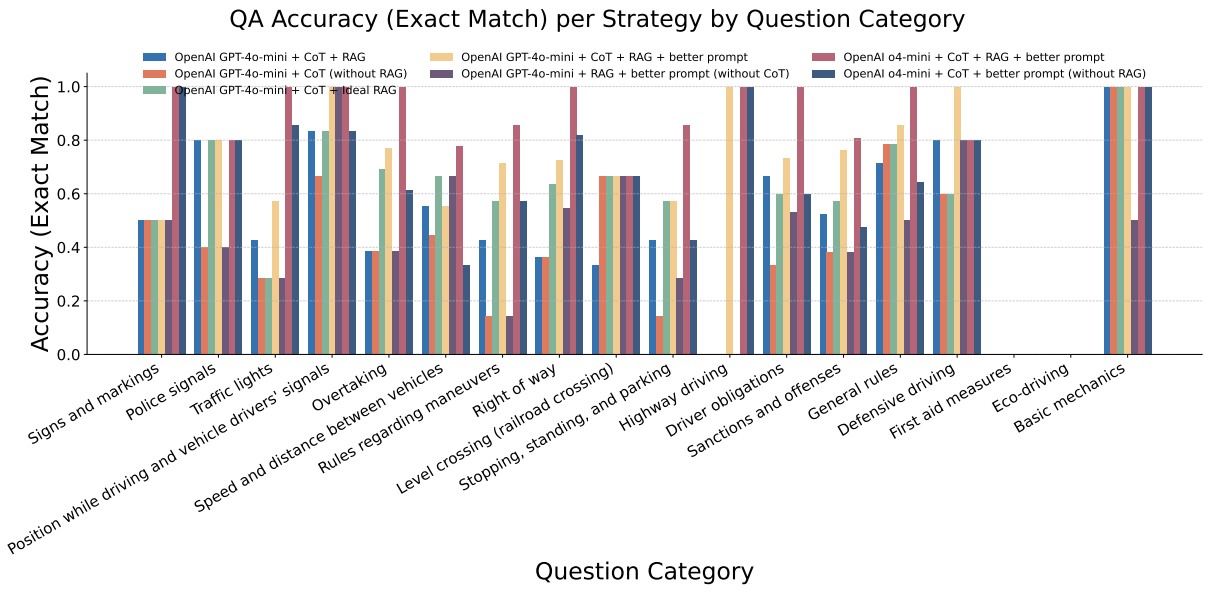


Figure 21: OpenAI models Exact Match score for QA per Strategy and Category on Split 1 Test

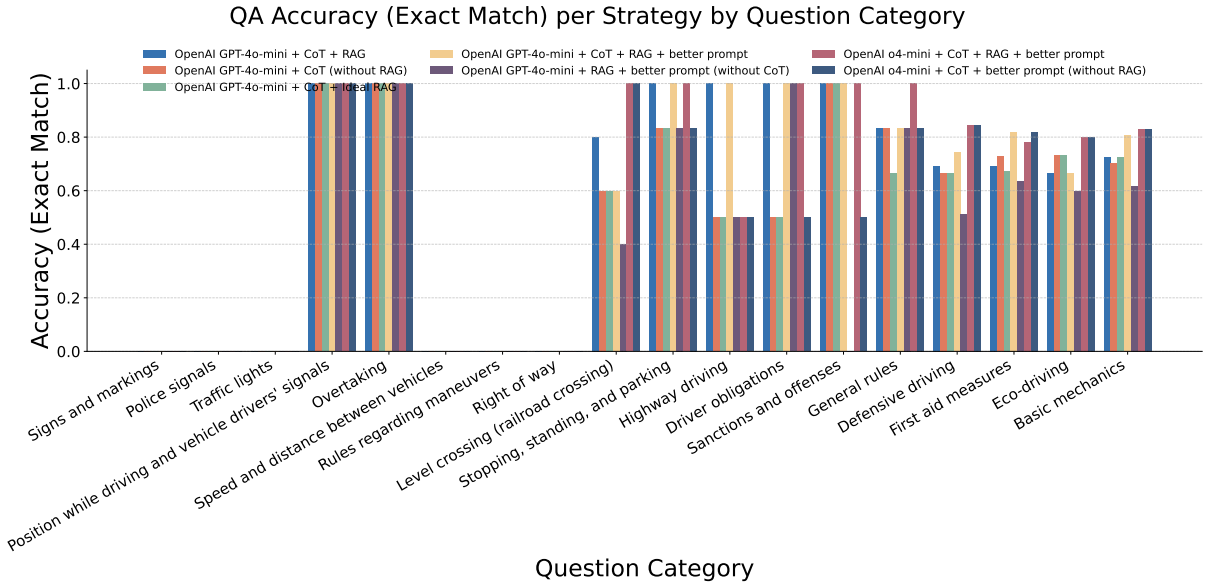


Figure 22: OpenAI models Exact Match score for QA per Strategy and Category on Split 2

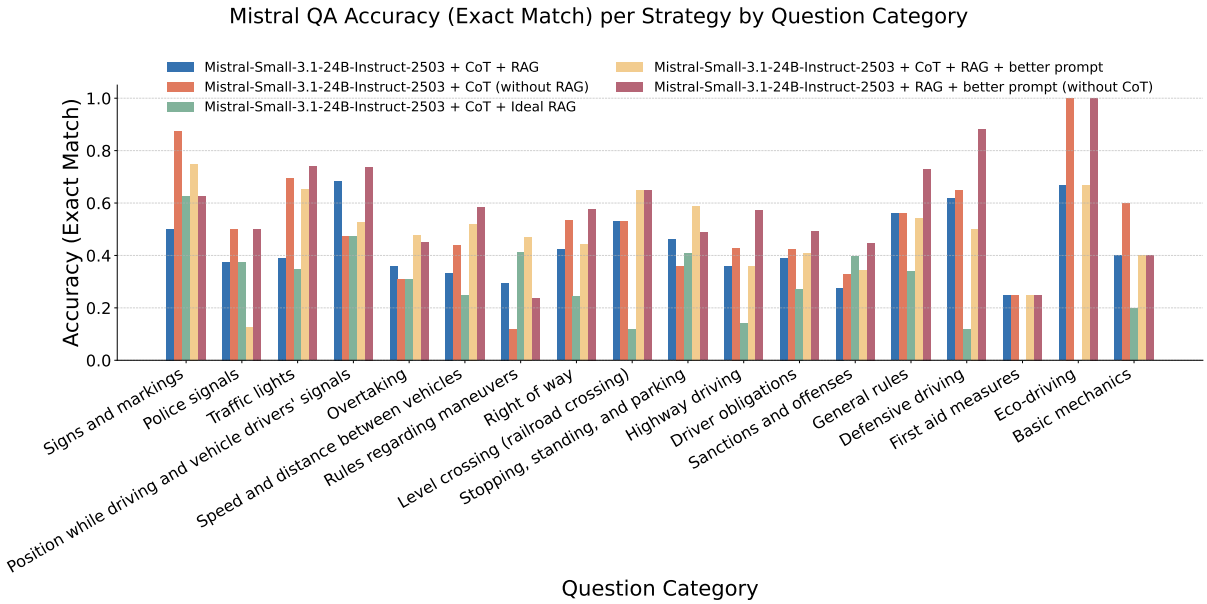


Figure 23: Mistral model Exact Match score for QA per Strategy and Category on Split 1 Train

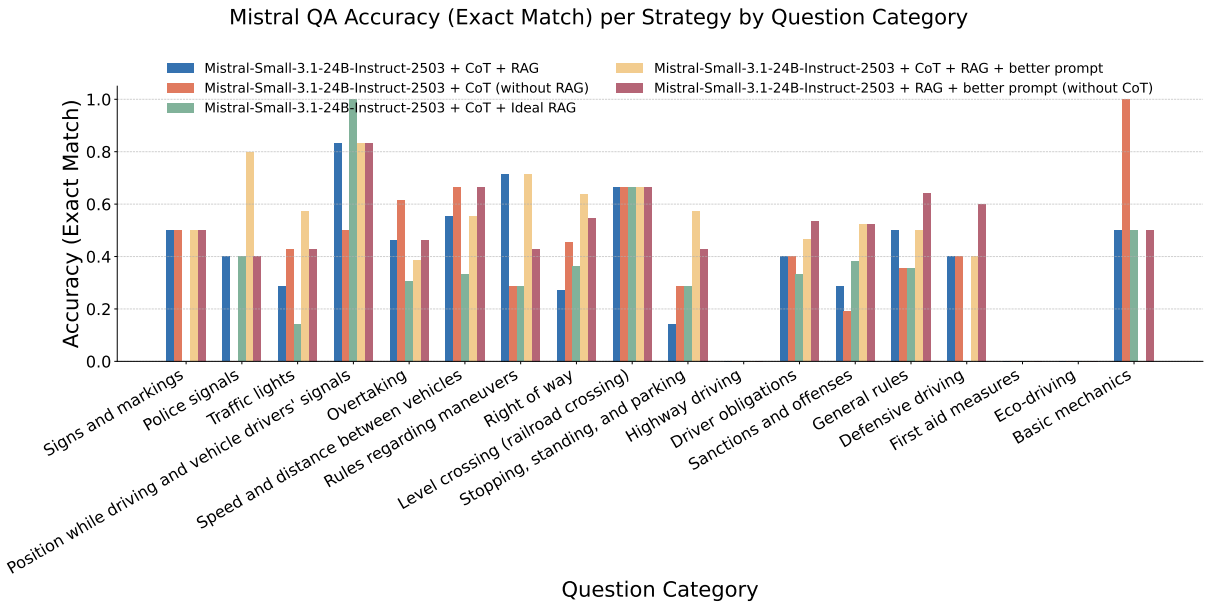


Figure 24: Mistral model Exact Match score for QA per Strategy and Category on Split 1 Test

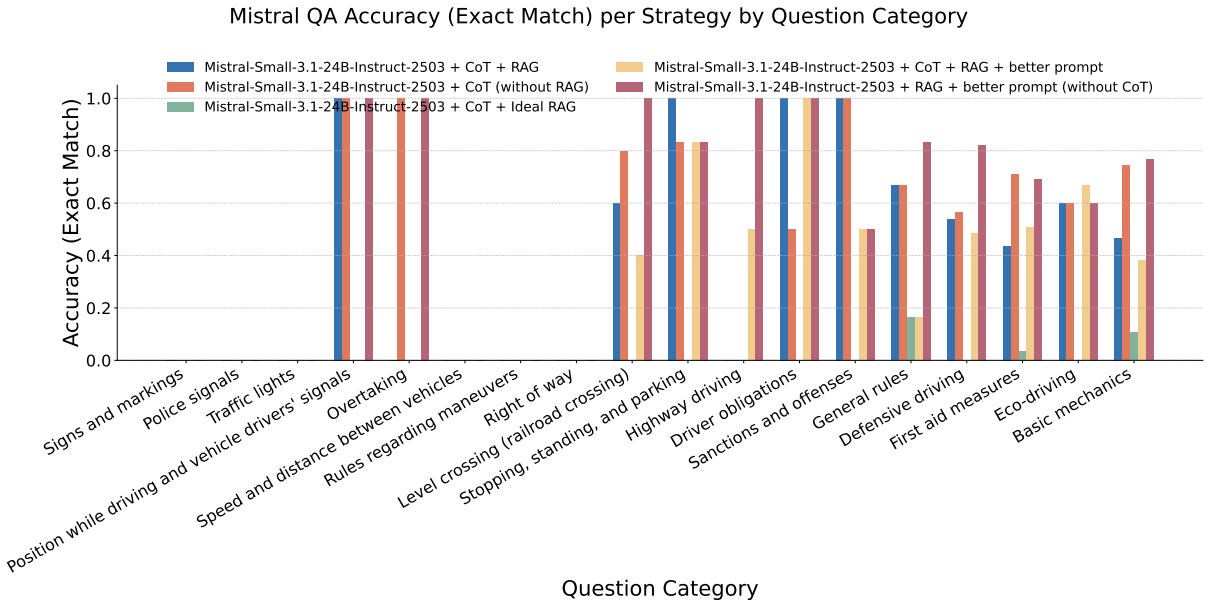


Figure 25: Mistral model Exact Match score for QA per Strategy and Category on Split 2

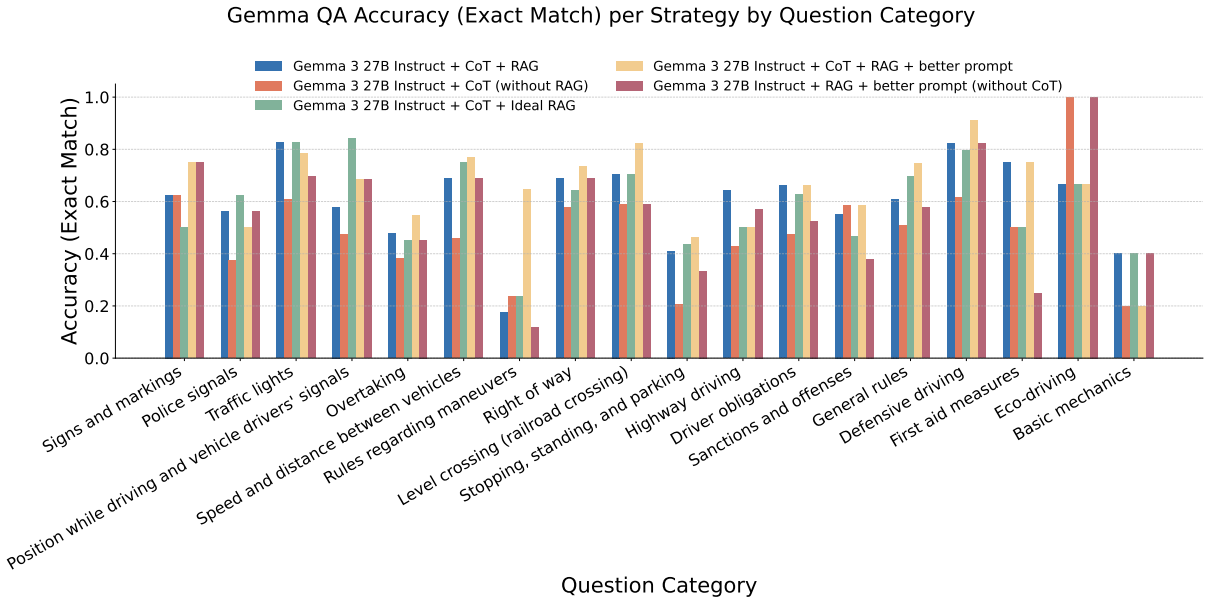


Figure 26: Gemma model Exact Match score for QA per Strategy and Category on Split 1 Train

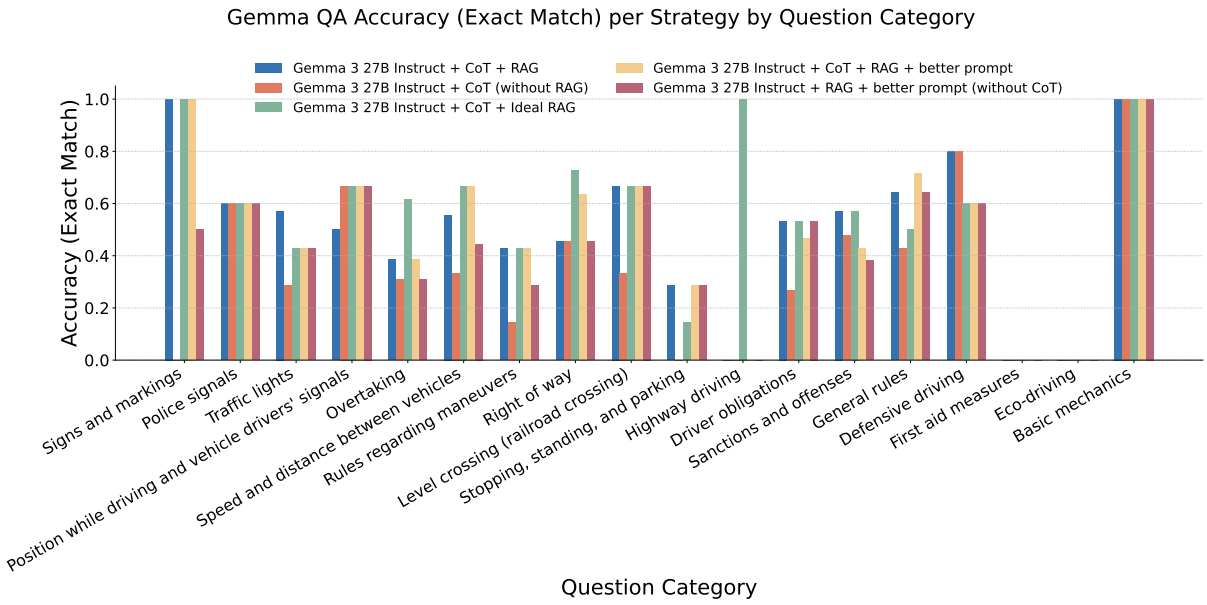


Figure 27: Gemma model Exact Match score for QA per Strategy and Category on Split 1 Test

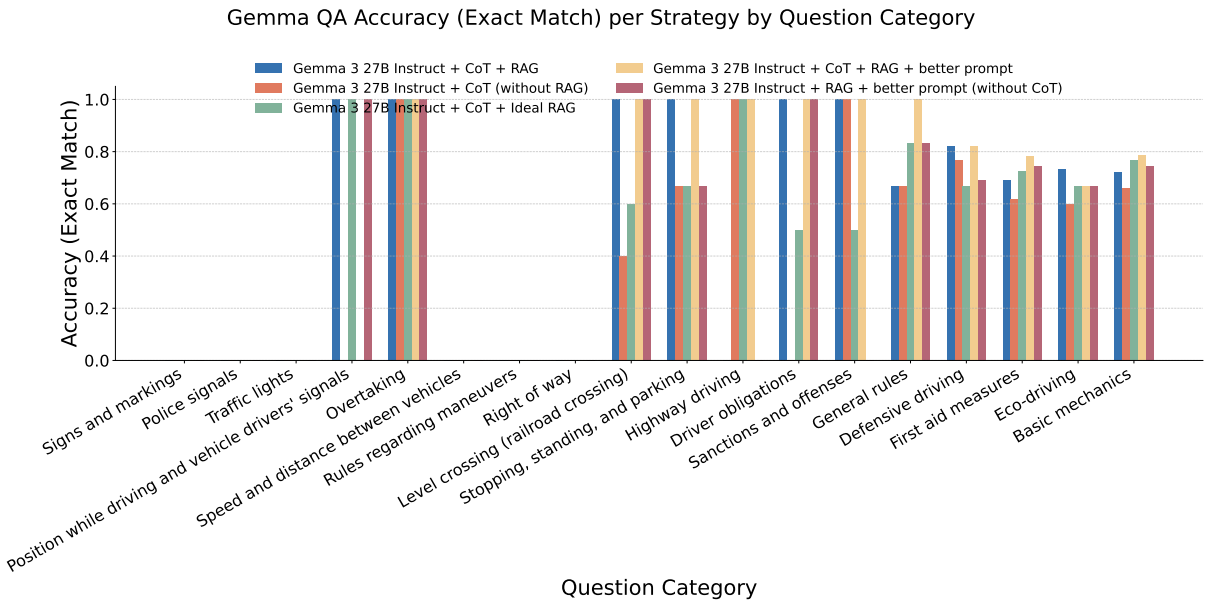


Figure 28: Gemma model Exact Match score for QA per Strategy and Category on Split 2

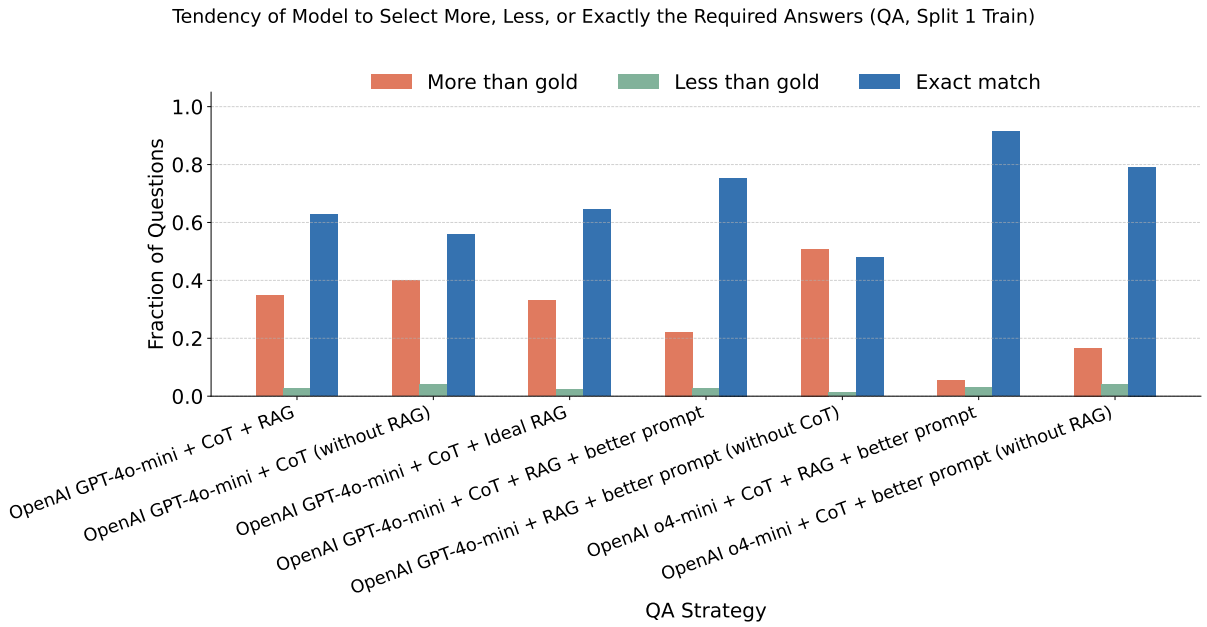


Figure 29: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 1 Train)

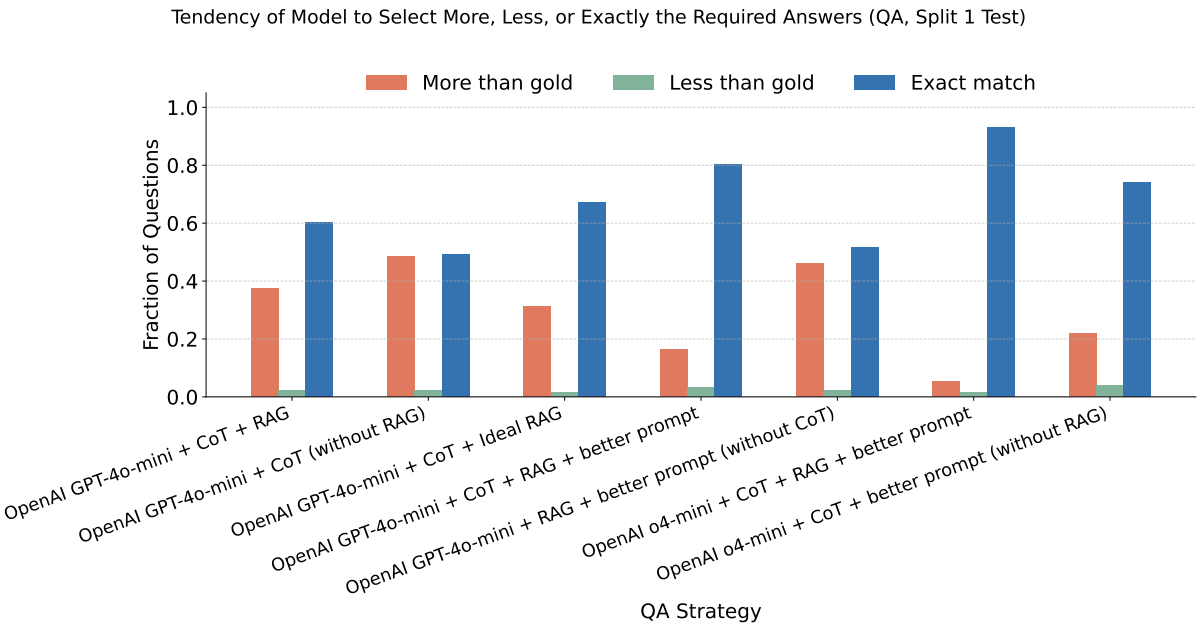


Figure 30: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 1 Test)

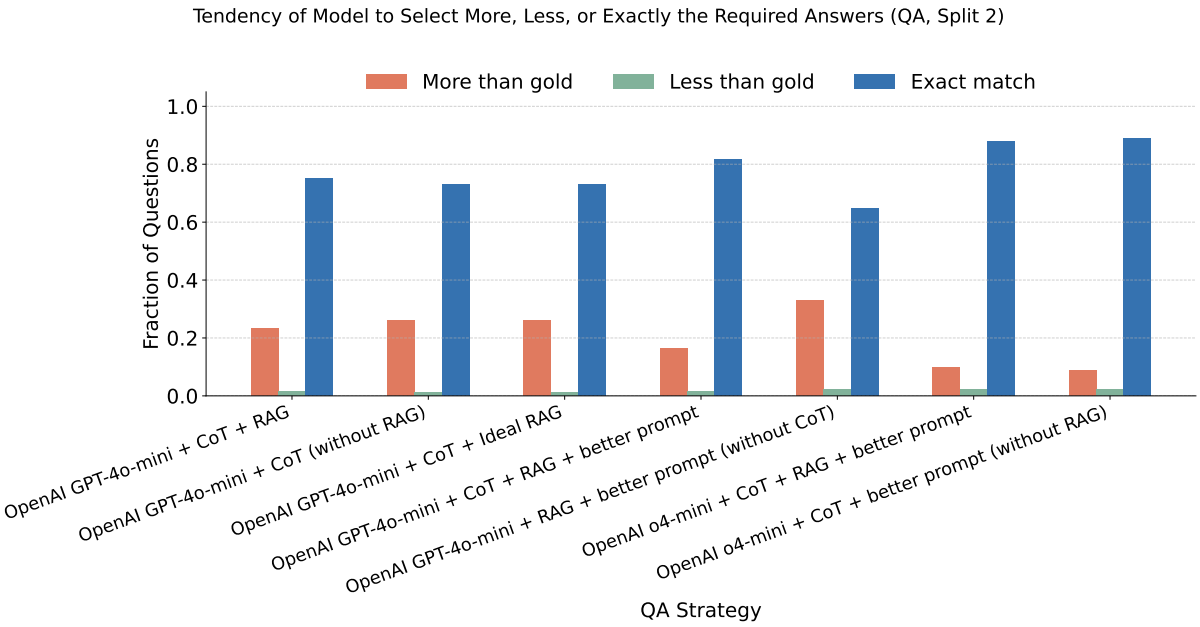


Figure 31: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 2)

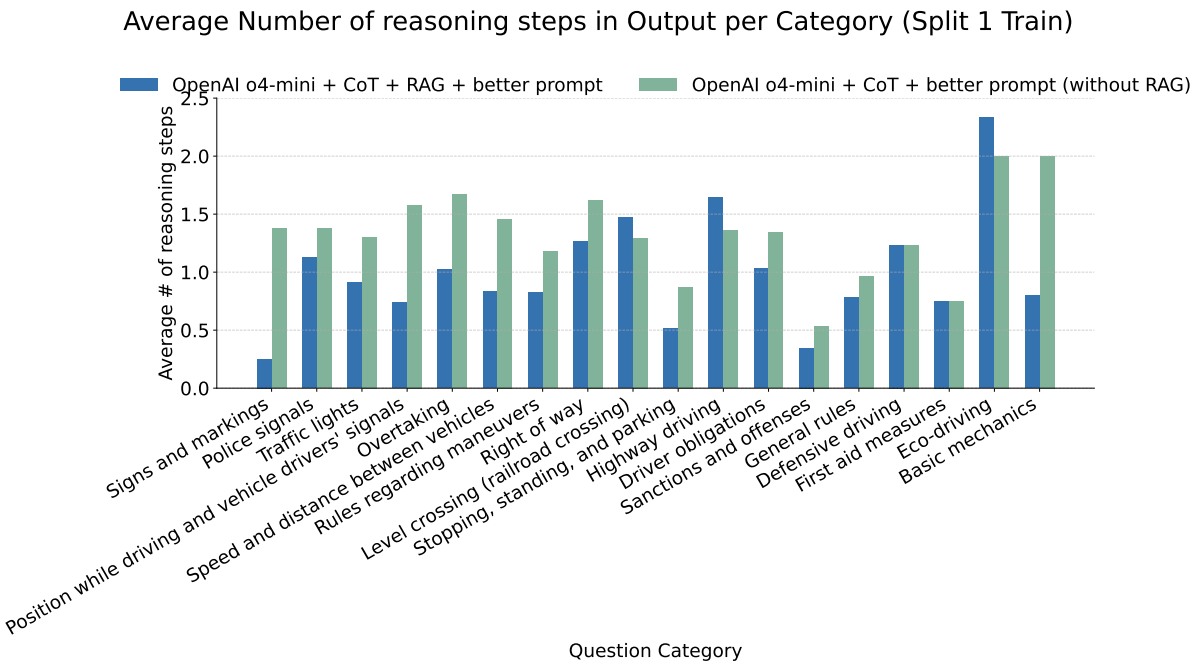


Figure 32: Average Number of reasoning steps in Output per Category (QA, Split 1 Train)

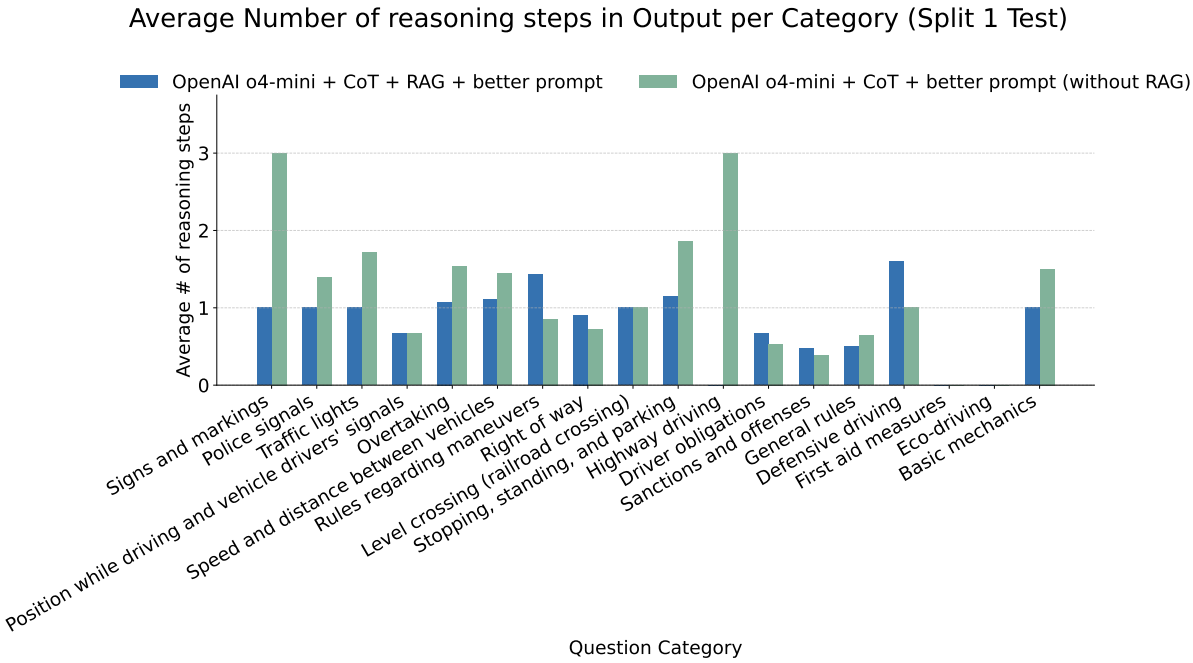


Figure 33: Average Number of reasoning steps in Output per Category (QA, Split 1 Test)

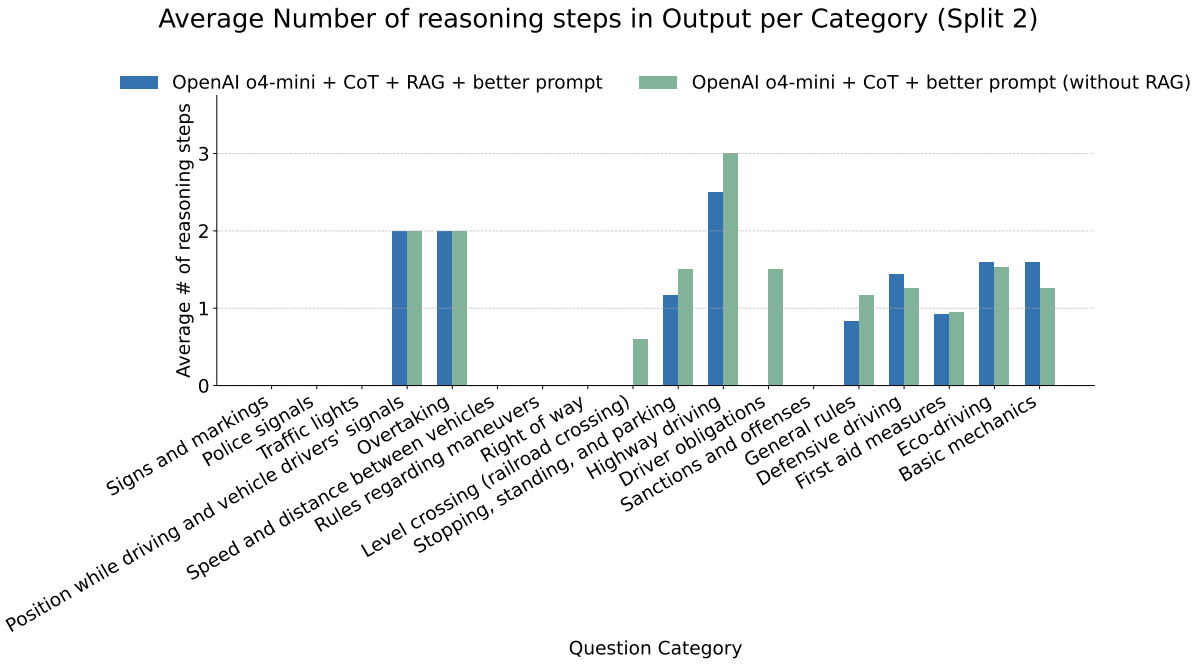


Figure 34: Average Number of reasoning steps in Output per Category (QA, Split 2)

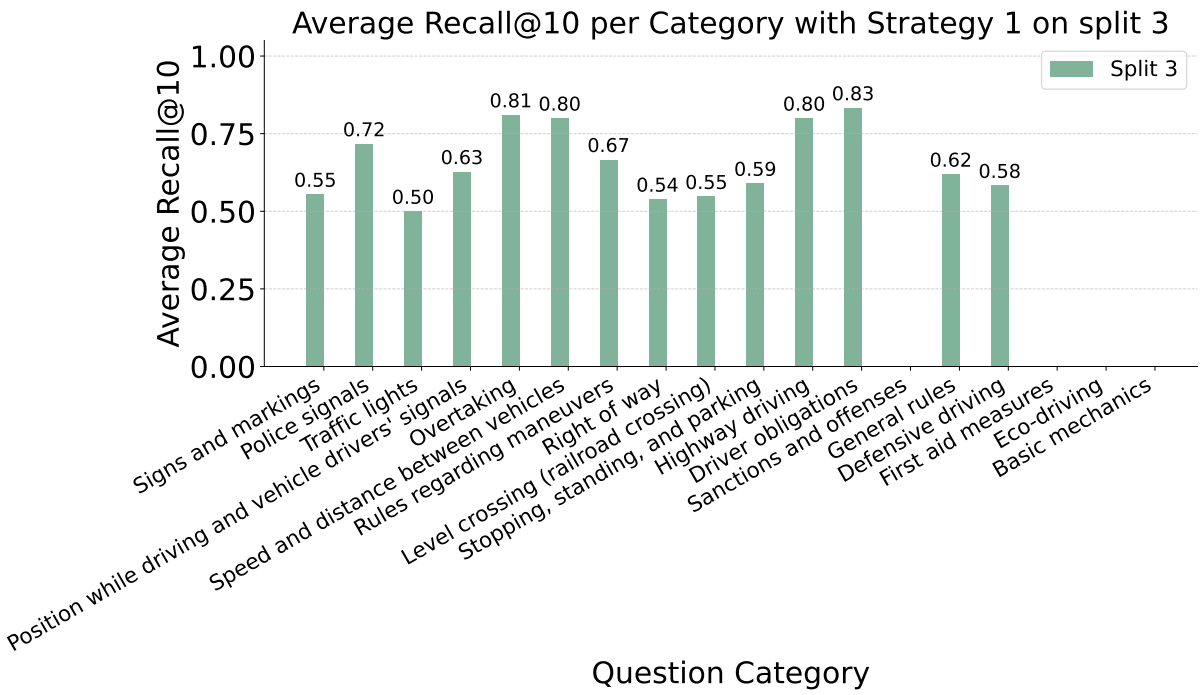


Figure 35: Recall@10 Analysis for VIR task with strategy 1 (laws)

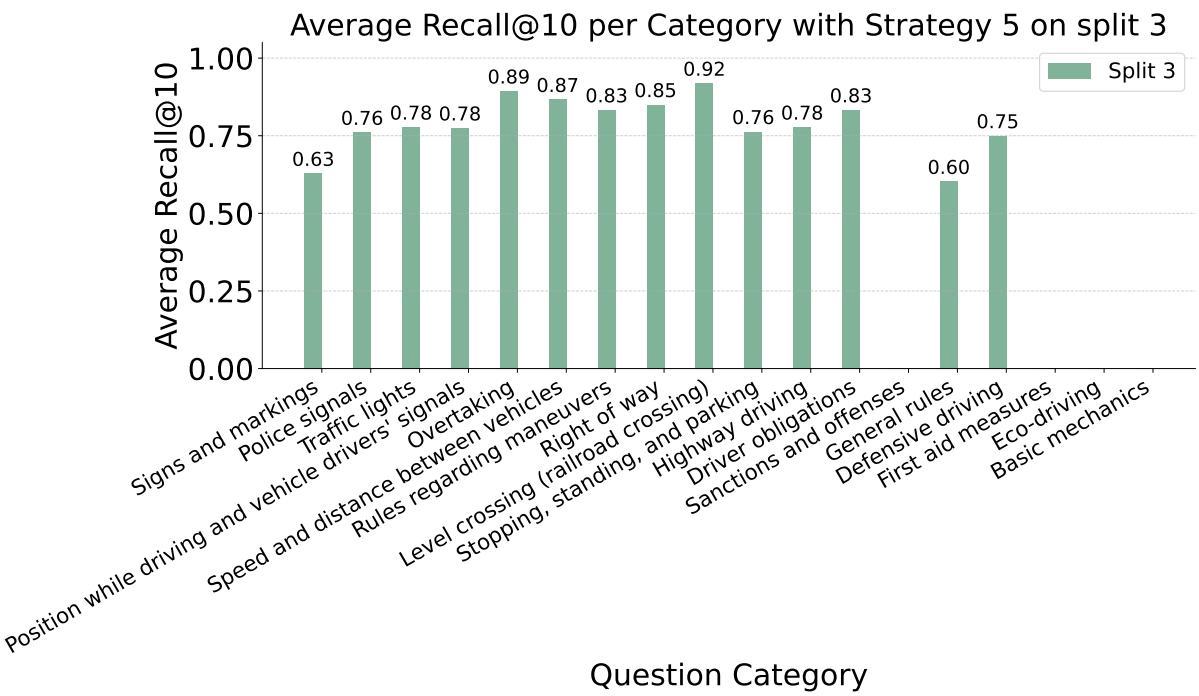


Figure 36: Recall@10 Analysis for VIR task with strategy 5 (laws)

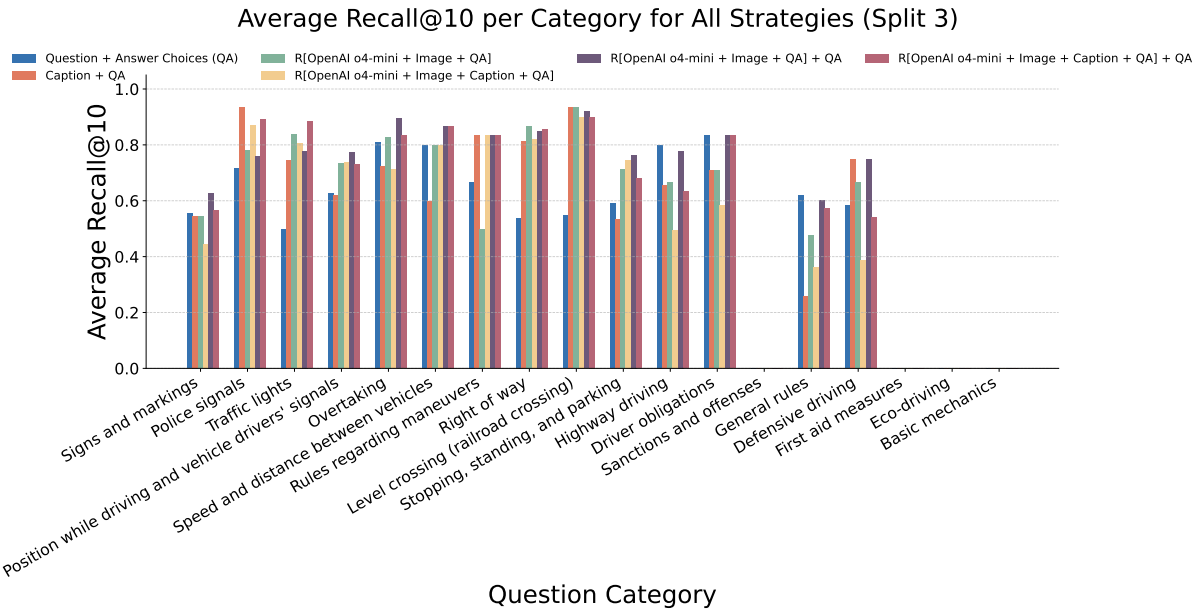


Figure 37: Recall@10 Analysis for VIR task with all strategies on Split 3 (laws)

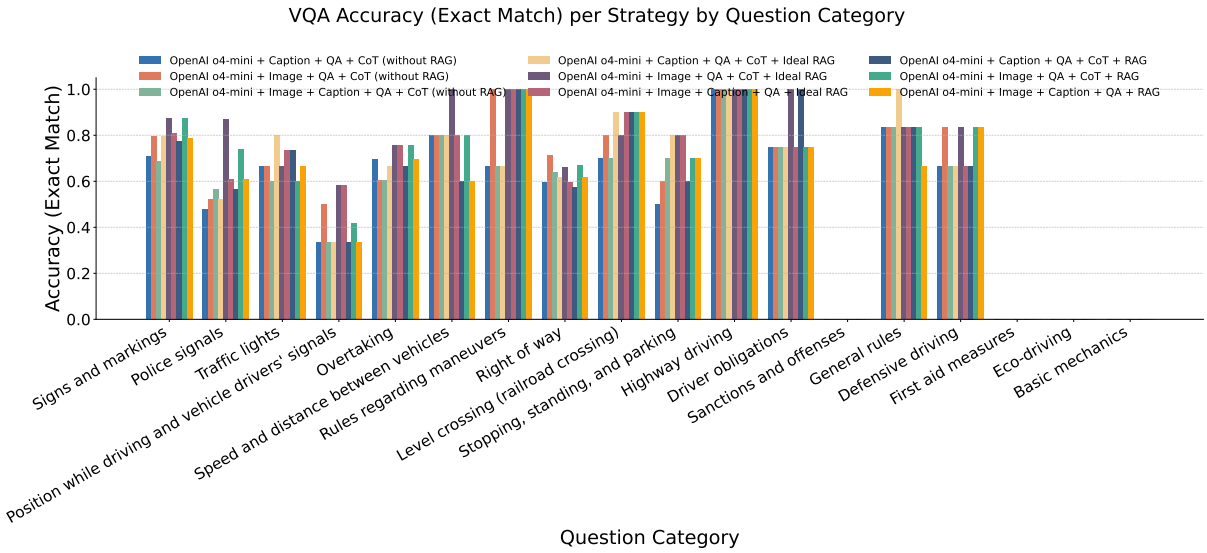


Figure 38: Exact Match score for VQA per Strategy and Category on Split 3

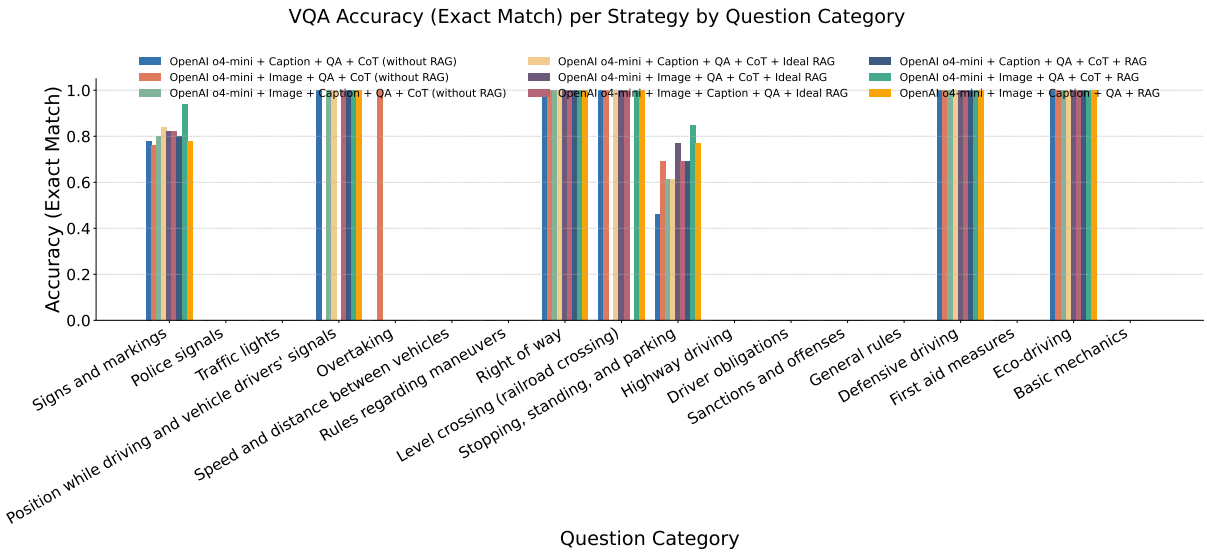


Figure 39: Exact Match score for VQA per Strategy and Category on Split 4

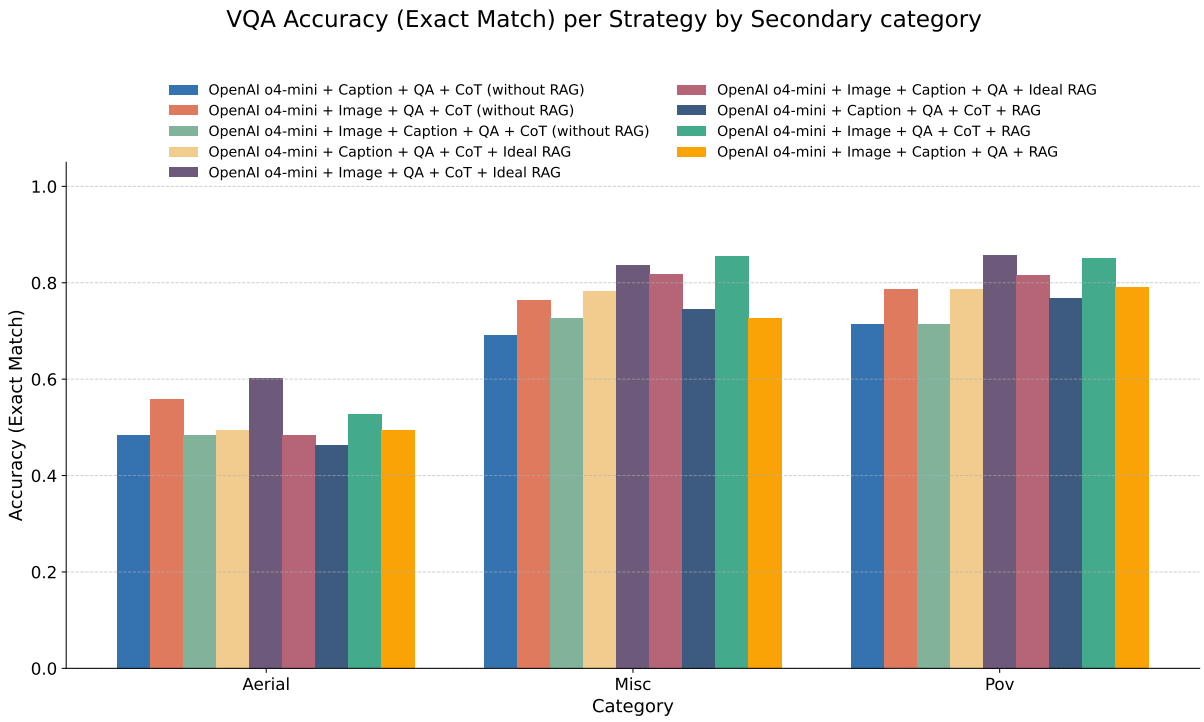


Figure 40: Exact Match score for VQA per Strategy and Secondary Category on Split 3

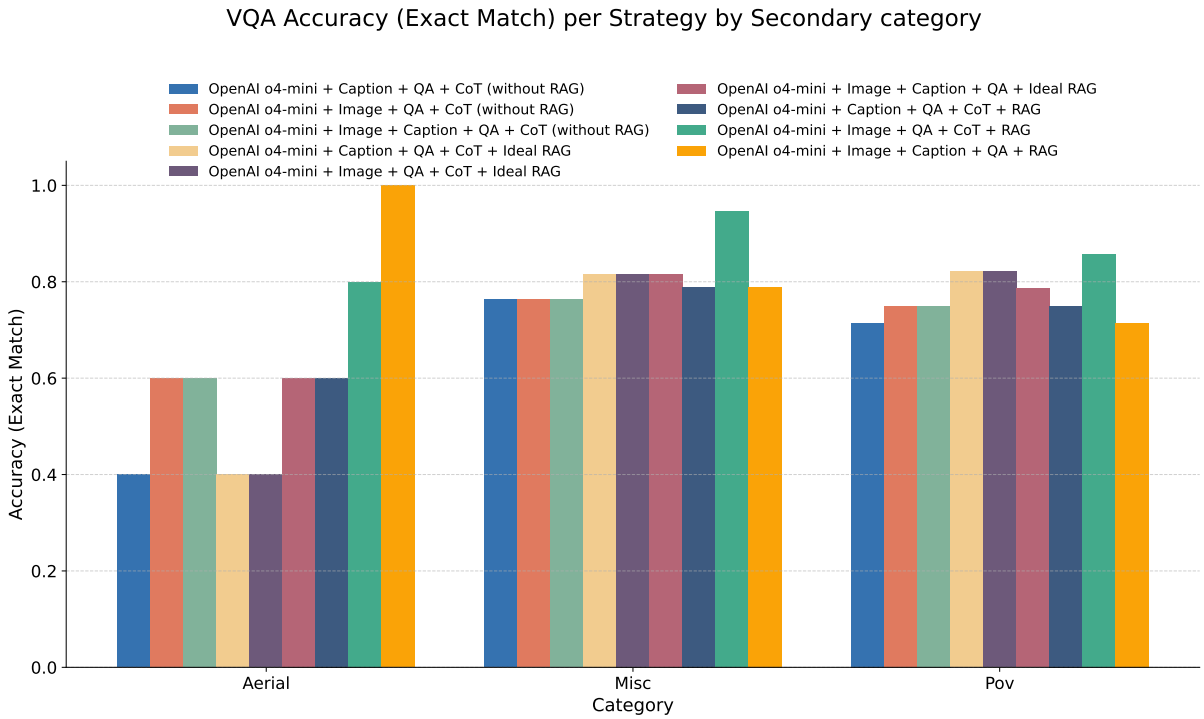


Figure 41: Exact Match score for VQA per Strategy and Secondary Category on Split 4

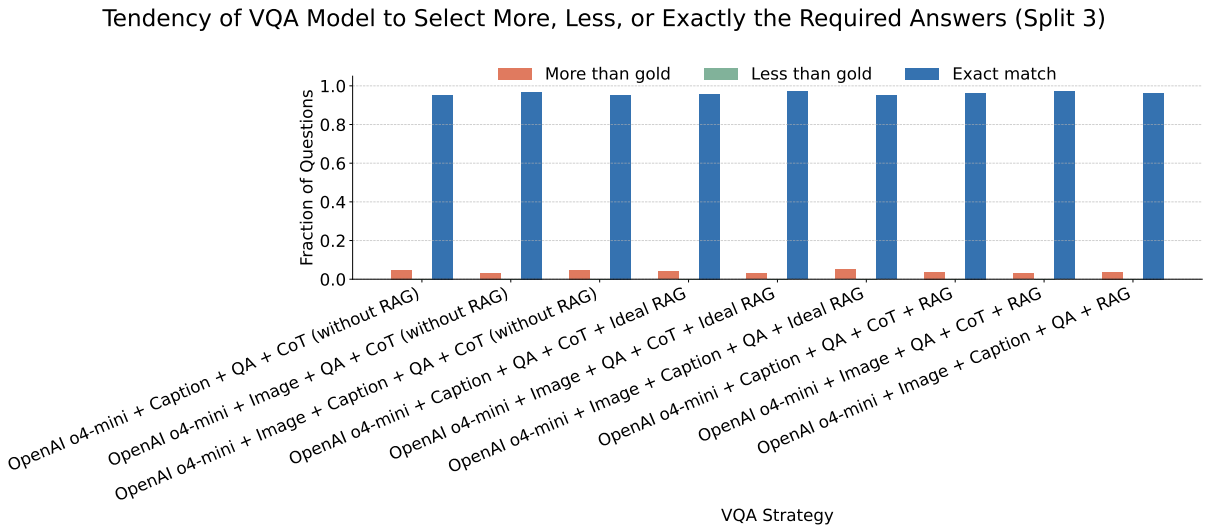


Figure 42: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (VQA, Split 3)

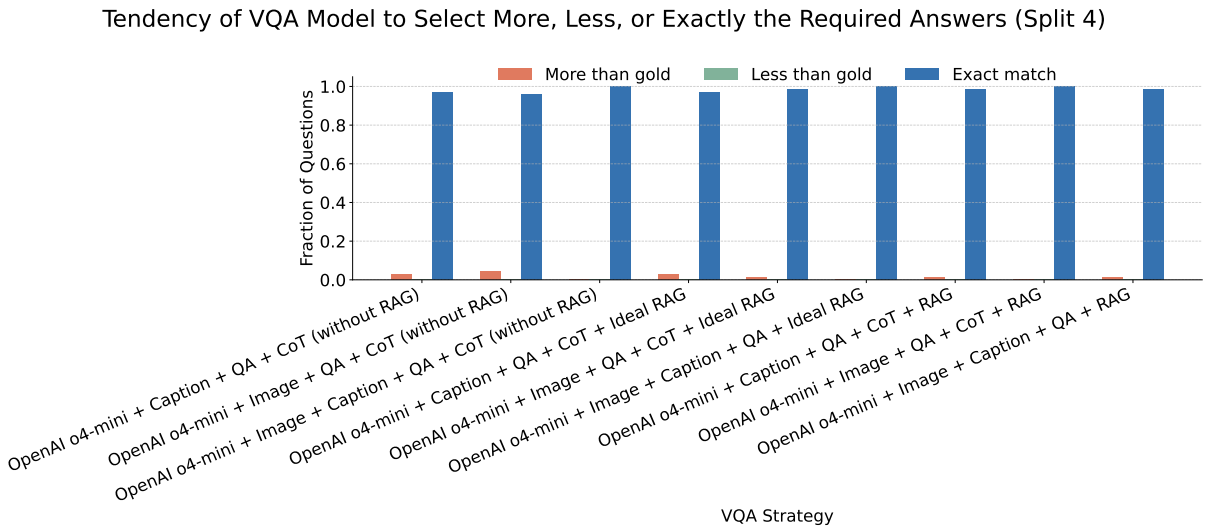


Figure 43: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (VQA, Split 4)

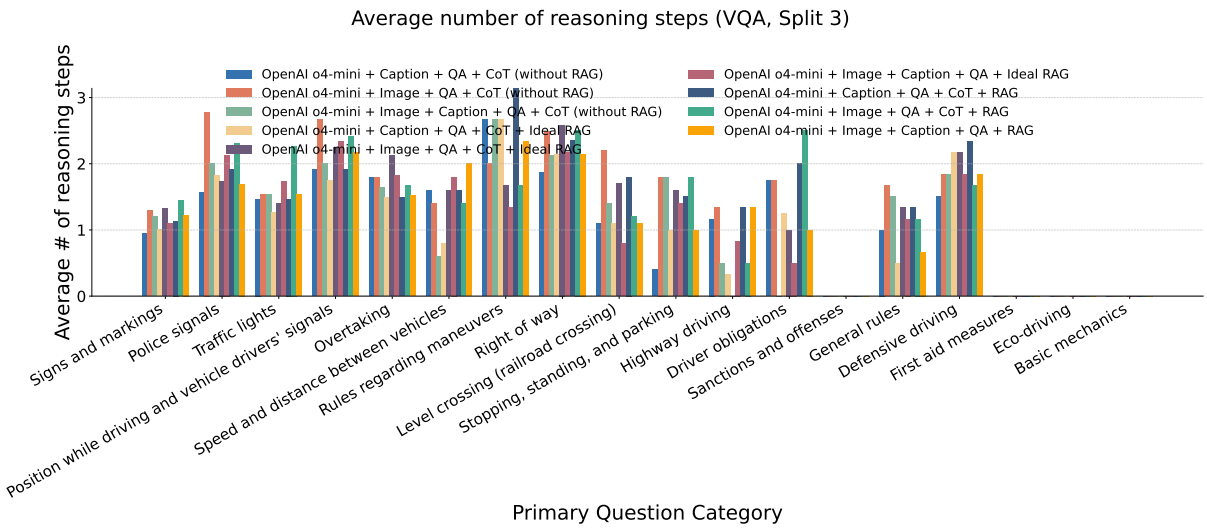


Figure 44: Average Number of reasoning steps in Output per Category (VQA, Split 3)

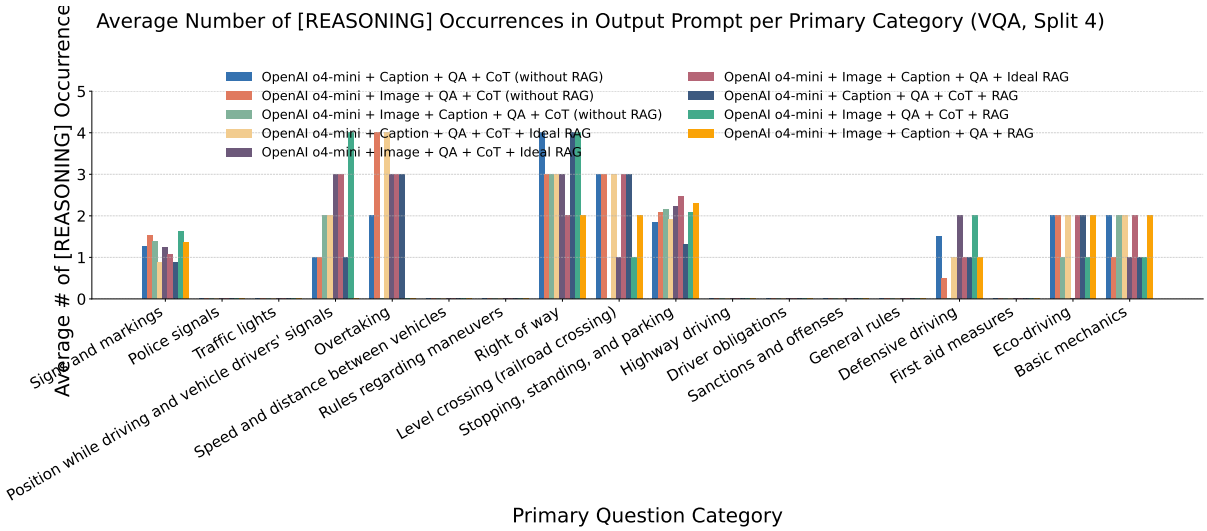


Figure 45: Average Number of reasoning steps in Output per Category (VQA, Split 4)

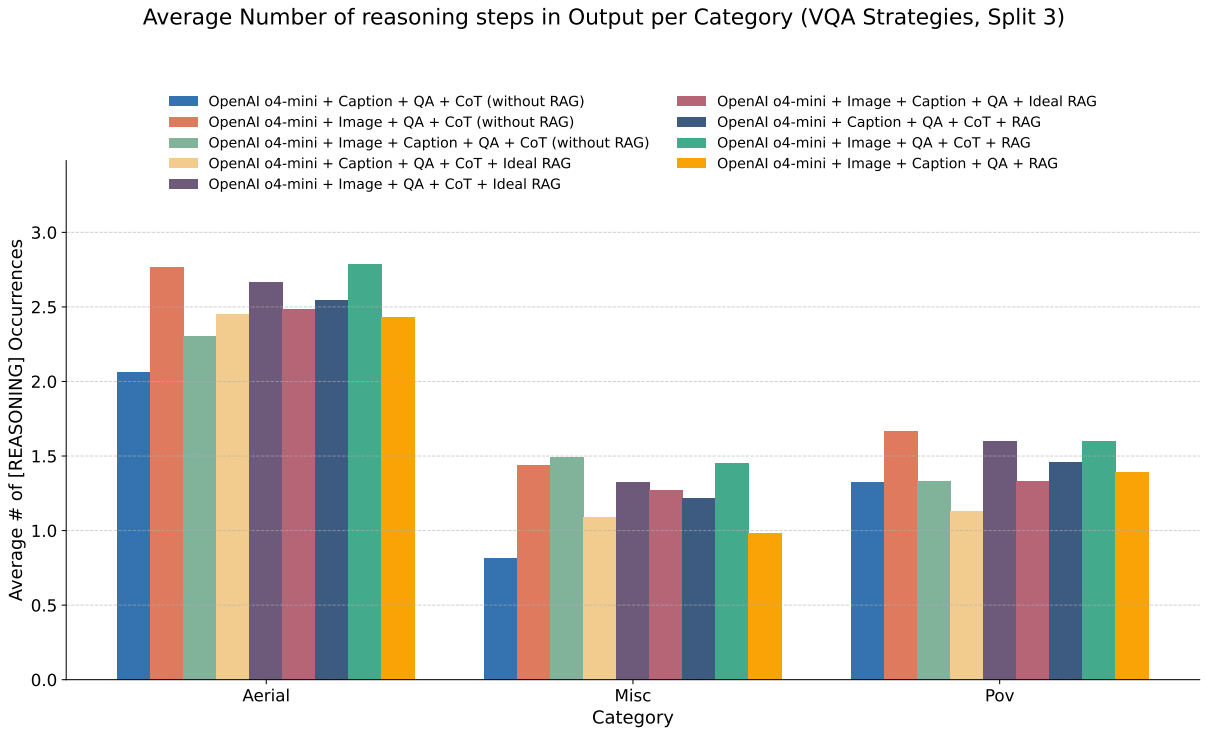


Figure 46: Average Number of reasoning steps in Output per Secondary Category (VQA, Split 3)

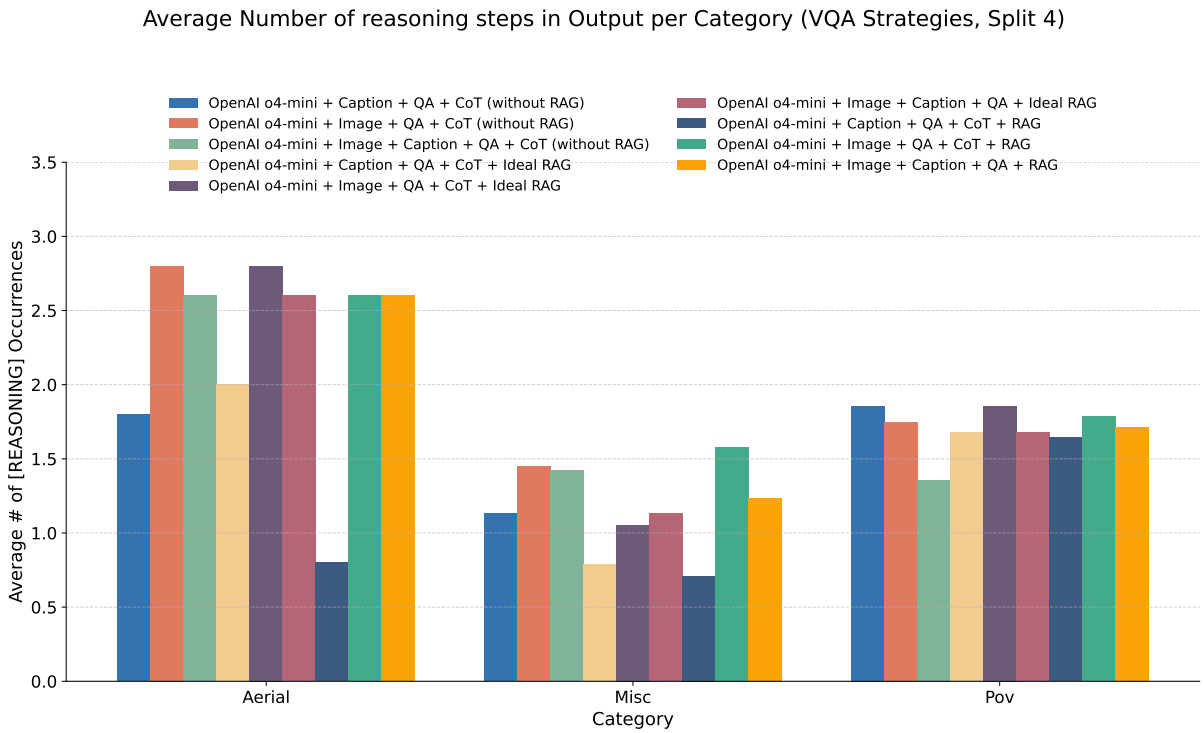


Figure 47: Average Number of reasoning steps in Output per Secondary Category (VQA, Split 4)